4

Itemresponstheorie

Het belangrijkste concept in de klassieke testtheorie is de betrouwbaarheid: daarmee wordt aangegeven in welke mate geobserveerde verschillen in toetsscores werkelijke verschillen tussen personen weerspiegelen. De definitie van de betrouwbaarheid steunt op de opsplits- baarheid van de variantie van de toetsscores X (zie hoofdstuk 3):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \tag{4.1}$$

of de variantie van de toetsscore, de totale variantie, is de som van de variantie van de ware scores plus de variantie van de meetfout. De betrouwbaarheid is dan per definitie de ver- houding tussen de variantie van de ware score en de totale variantie:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XX'}. \tag{4.2}$$

Het rechterlid van (4.2) geeft aan hoe die betrouwbaarheid kan worden vastgesteld, namelijk als de correlatie tussen X en een parallelvorm X'. Indien we (4.2) wat nader onderzoeken dan duiken er twee problemen op waarvoor niet zo snel een oplossing gevonden is.

Het eerste probleem betreft het gebruik van spreidingsmaten, zoals de variantie, die altijd naar een verdeling of een populatie verwijzen. Hoewel dit in (4.2) niet uitdrukkelijk gezegd wordt, is de referentie naar een of andere populatie impliciet aanwezig, en dit impliceert weer dat de betrouwbaarheid van een toets een eigenschap is die niet alleen de toets karakteriseert, maar de toets in de populatie. Het niet expliciteren van die betrekkelijkheid, wat in de praktijk nogal eens voorkomt, dekt het

probleem misschien toe, maar lost het zeker niet op. Een mededeling zoals "de betrouwbaarheid van X is 0.8" is dus zinloos als men er zich niet van verzekert dat spreker en aangesprokene dezelfde populatie in gedachten hebben.

Het tweede probleem is dat de ware score T toetsspecifiek is: de intuïtieve betekenis van de ware score is de gemiddelde score die een persoon behaalt wanneer de toets X een zeer groot aantal keer onder dezelfde omstandigheden wordt afgenomen. Het is daarbij irrelevant of dit al dan niet praktisch realiseerbaar is. De belangrijke vraag is echter of het kennen of schatten van deze ware toetsscore op zichzelf een belangrijke aangelegenheid is. In theoretisch onderzoek en in toepassingen zal men toch eerder tot het standpunt neigen dat een toetsscore iets dient te onthullen over een meer abstracte entiteit, een vaardigheid, een geschiktheid of een attitude, waarbij de items die men in de toets gebruikt in principe zouden kunnen worden vervangen door andere items. De belangrijke vraag is dus of de ware toetsscore, die samenhangt met een specifieke toets, iets kan zeggen over een meer abstracte, onderliggende vaardigheid. Dit resulteert in een aantal vragen waarop de klassieke testtheorie geen afdoend antwoord kan bieden.

Een toets bestaat uit een aantal onderdelen of items. Hoe kan een toetsconstructeur weten of het zinvol is bepaalde items samen in dezelfde toets op te nemen? Immers, als de toetsscore een indicator is van de mate waarin een theoretisch concept aanwezig is of beheerst wordt, dient elk item dat in de toets wordt opgenomen relevant te zijn voor dit concept, dat wil zeggen de toets moet homogeen zijn met betrekking tot dit concept. Nu is het natuurlijk niet zo dat professioneel gemaakte toetsen een willekeurig allegaartje van items zijn. De toetsconstructeur gebruikt wel degelijk theoretische kennis om tot een verantwoorde keuze van items te komen. Het belangrijke punt is echter dat de klassieke testtheorie, als statistische theorie, geen middelen aanbiedt aan de hand waarvan duidelijk kan beslist worden of deze homogeniteit in conceptuele relevantie al dan niet bereikt is. Het beste wat de klassieke theorie kan bieden is een index van interne consistentie, de KR-20 bijvoorbeeld, maar zulke indices hebben een dubbelzinnige betekenis. Indien ze hoog zijn, waarbij de vraag wat hoog is een nieuw probleem oproept, dan wijst dit op homogeniteit en grote betrouwbaarheid. Echter, indien de KR-20 laag is, wijst dit op een gebrek aan homogeniteit of betrouwbaarheid of beide, en uit de waarde van de KR-20 valt niet af te leiden wat er nu precies het geval is.

De tweede vraag betreft de scoringsregel. In de klassieke testtheorie wordt de toetsscore bij dichotome items meestal gedefinieerd als het aantal items juist, ook wel aangeduid als ruwe somscore. Hoewel deze definitie voor de hand liggend kan lijken, is ze in principe willekeurig. Er zijn andere scoreregels denkbaar die in bepaalde omstandigheden veel zinvoller kunnen zijn. De klassieke benadering bevat echter geen

theorie waaruit de superioriteit van de gewone somscoreregel of welke regel dan ook volgt.

De derde vraag, die binnen de klassieke testtheorie in principe onoplosbaar is, is de volgende. Een steekproef van kinderen, aselect getrokken uit een goed gedefinieerde populatie, wordt op tijdstip t_1 gemeten met een toets X_1 en op tijdstip t_2 met een toets X_2, waarbij het de bedoeling is te schatten of de gemiddelde vaardigheid in de populatie veranderd is in het interval $(t_1 - t_2)$. Indien X_1 niet identiek is aan X_2 treedt er een dubbel probleem op. Indien het gemiddelde op X_2 groter is dan het gemiddelde op X_1 zou het verschil te wijten kunnen zijn aan het feit dat X_2 gemakkelijker is dan X_1, of aan het feit dat de gemiddelde vaardigheid inderdaad is toegenomen, of aan beide. Om de verklaring van een gemakkelijker toets uit te sluiten dienen dus speciale maatregelen genomen te worden, bijvoorbeeld het afnemen van toets X_2 op tijdstip t_1 bij een onafhankelijke steekproef uit dezelfde populatie, zodanig dat X_1 en X_2 kunnen geëquivaleerd worden (zie hoofdstuk 8). Equivaleren is echter een puur technische ingreep, en is zeker geen oplossing voor het tweede, veel fundamenteler probleem: hoe kan gegarandeerd worden dat X_1 en X_2 inderdaad hetzelfde concept meten. Indien men op dit probleem geen afdoende antwoord kan geven staat men weerloos tegen de aantijging dat bovengenoemde vergelijking het vergelijken is van appels met peren, en dus zinloos.

In de moderne testtheorie wordt aan de eerdergenoemde twee problemen van de klassieke testtheorie, te weten de populatie-afhankelijkheid en de toetsspecificiteit van de score, tegemoet gekomen. De theorie wordt ontwikkeld zonder enige referentie aan een of andere populatie, hoewel we verderop zullen zien dat in sommige omstandigheden dit populatiebegrip weer zal opduiken. Bovendien staat in die theorie niet de toetsscore centraal, maar het item en het antwoord op het item. Dit verklaart meteen ook de naam van deze theorie: itemresponstheorie (IRT). Hiervoor hebben we gezegd dat de ware score T van een persoon in principe observeerbaar is door de scores van een groot aantal toetsafnames te middelen. De IRT hanteert een begrip dat men losjes zou kunnen omschrijven als de te meten vaardigheid, dat in principe niet observeerbaar is. Om deze principiële onobserveerbaarheid aan te duiden gebruikt men de term latent, en het begrip vaardigheid wordt soms vervangen door de meer neutrale term trek. Een equivalente doch verouderde benaming voor IRT is dan ook latente-trektheorie (in het Engels: latent trait theory).

Een IRT is een geheel van uitspraken over de samenhang tussen de latente trek en het antwoordgedrag op een verzameling items. De conceptuele homogeniteit waarover hierboven werd gesproken is niets anders dan deze samenhang. In de mate dat deze samenhang duidelijk gedefinieerd is, weten we ook wat precies met homogeniteit wordt

bedoeld. In paragraaf 4.1 wordt een algemene inleiding van deze theorie gegeven aan de hand van één speciaal geval, het Raschmodel.

De uitspraken in zo'n theorie zijn meestal niet heel specifiek: de voorspellingen over het gedrag hangen af van kenmerken van de items en van de personen. Deze kenmerken worden meestal gekwantificeerd als kengetallen of parameters, en de waarden van deze parameters zijn in de regel niet bekend. Een belangrijk probleem in de IRT is dan ook het schatten van deze parameters en het geven van een aanduiding van de nauwkeurigheid waarmee deze parameters kunnen worden geschat. De schattingsproblematiek wordt behandeld in paragraaf 4.2.

Een theorie is alleen die naam waardig indien ze gefalsificeerd kan worden. In paragraaf 4.3 worden methoden besproken waarmee kan worden nagegaan of de predicties over het gedrag die uit de theorie volgen wel met de werkelijkheid overeenkomen. Deze methoden steunen sterk op de statistische theorie, en nemen meestal de vorm aan van formele statistische toetsen waarbij het gehanteerde model de status van nulhypothese krijgt.

Paragraaf 4.4 bevat een technische uiteenzetting van de werkwijze bij parameterschattingen en modeltoetsen indien de data verzameld zijn in een onvolledig design.

Men kan zich natuurlijk gaan afvragen waar de meetprocedure zelf blijft. De bedoeling van het meten is het toekennen van een getal aan een persoon op zodanige manier dat de grootte van het getal ook de mate van zijn vaardigheid uitdrukt. Het is kenmerkend voor de literatuur in IRT dat de eerste en meeste aandacht gaat naar het zorgvuldig opbouwen en toetsen van de theorie, en dat de meetprocedures zelf veel minder aandacht krijgen. Niettemin is de meetprocedure zelf belangrijk en een aantal subtiele problemen in verband hiermee verdienen meer aandacht dan ze doorgaans in de literatuur krijgen. Dit is het onderwerp van paragraaf 4.5.

4.1 Begrippen en algemene theorie

Centraal in de IRT staat het begrip latente variabele. Hoewel er verschillende opvattingen zijn over de status van deze variabele, zullen we ons hier beperken tot één geval, namelijk waar het domein van de latente variabele de reële as is. Elke persoon in een populatie kan afgebeeld worden als een punt van de reële as, of wat equivalent hiermee is, aan elke persoon kan een getal worden toegevoegd dat een uitdrukking is van de mate waarin die persoon over de vaardigheid beschikt. Aan die latente variabele geen inhoud toegeschreven, het is dus een abstracte variabele, die we verder dan ook

met het algemeen symbool θ zullen aanduiden. De getalswaarde die aan persoon *v* is toegekend duiden we aan als θ_v.

Merk op dat de waarde van θ niet begrensd is: $-\infty < \theta < \infty$. Om iets te kunnen zeggen over de θ-waarde van een persoon veronderstelt men dat de antwoorden op bepaalde items enige indicatie geven over de vaardigheid. Bijvoorbeeld door een uitspraak als: "een correct antwoord op dit item duidt op een grotere vaardigheid dan een fout antwoord". Met zo'n vage uitspraak kan natuurlijk niet veel gedaan worden. In de IRT staat het expliciet maken van het verband tussen de latente variabele θ en de itemantwoorden dan ook centraal.

Eerst een definitie. Met X_i duiden we het antwoord aan op item *i*, en voorlopig gaan we ervan uit dat X_i dichotoom is, met waarden toegekend volgens onderstaande regel:

$$X_i = \begin{cases} 1 \text{ indien het antwoord op item } i \text{ correct is,} \\ \\ 0 \text{ indien het antwoord op item } i \text{ fout is.} \end{cases}$$

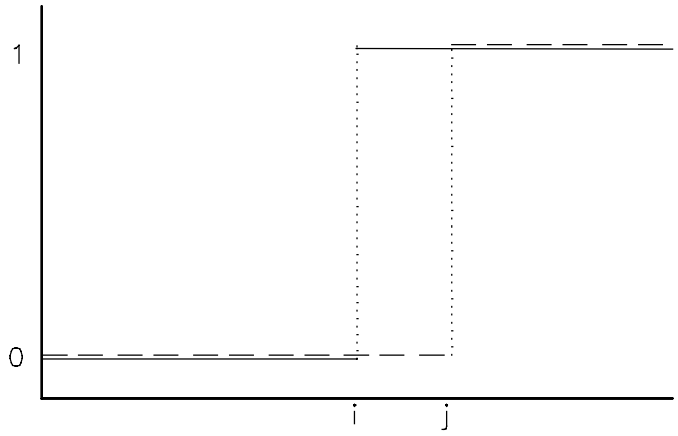
Centraal in de IRT is de aanname dat het antwoord op een item nooit volledig vastligt, hoe groot of hoe klein de vaardigheid van de persoon die het item beantwoordt ook is. Daarom wordt met kansen gewerkt, en de variabele X_i is een toevalsvariabele. De itemresponsfunctie drukt uit hoe groot de kans is dat het item juist wordt beantwoord als functie van de vaardigheid. Deze functie wordt aangeduid met het symbool $f_i(\theta)$. Dus,

$$f_i(\theta) = P(X_i=1|\theta) \qquad (4.3)$$

of, de itemresponsfunctie is de conditionele kans op een juist antwoord gegeven de waarde van θ. Formule (4.3) is nog geen theorie; zij is eigenlijk niets meer dan een conventie over de notatie. We schrijven kortheidshalve het linkerlid op, als we het rechterlid bedoelen. Om een echte theorie te maken zullen we de functie moeten specificeren, dat wil zeggen we moeten het verloop ervan beschrijven en er de eigenschappen van vastleggen. Omdat we later mathematische manipulaties met die functie zullen moeten uitvoeren, zullen we eisen dat ze niet te gek is en dat ze geloofwaardig is. Voor een goed begrip van de theorie beginnen we echter met een niet-geloofwaardige functie, die als volgt geconstrueerd wordt. Voor een item *i* veronderstelt men dat er een bepaalde hoeveelheid vaardigheid nodig is om een correct antwoord te produceren. Iemand die over minder vaardigheid beschikt zal nooit een correct antwoord geven, de kans op een correct antwoord is 0, terwijl iemand met meer

vaardigheid het item altijd juist beantwoordt, dat wil zeggen met kans 1. De grafiek van de itemresponsfunctie is weergegeven in figuur 4.1. Merk op dat de grafiek van de functie een sprong maakt op de plaats *i*. In dezelfde figuur is ook de plaats aangegeven voor een moeilijker item *j*. Dit item is moeilijker dan item *i*, omdat de minimale vaardigheid vereist voor een correct antwoord op item *j* groter is dan voor item *i*.

Deze theorie ziet er misschien aantrekkelijk uit, want ze impliceert het principe: wie een moeilijk item *(j)* juist beantwoordt, geeft ook een juist antwoord op een gemakkelijker item *(i)*. Een verzameling items, waarbij bovenstaande uitspraak geldig is voor alle paren wordt een Guttman-schaal genoemd, naar een van de grondleggers van de moderne testtheorie (Guttman, 1950). Deze theorie is echter niet erg geloofwaardig, omdat het in de praktijk bijna nooit voorkomt dat er in de steekproef niemand is die dit principe schendt. Eén inbreuk op dit principe is voldoende om de theorie te verwerpen. Uit inspectie van figuur 4.1 konden we eigenlijk al dit soort moeilijkheden verwachten. Omdat de kans op een juist antwoord altijd precies 0 of 1 is, leggen we de waarde van X_i volledig vast als we θ kennen, en in de praktijk kunnen we daarvoor gestraft worden. Dergelijke modellen noemt men deterministisch. In de IRT werkt men meestal met itemresponsfuncties die nooit exact de waarde 0 of 1 aannemen. Een andere eigenschap die de functies in figuur 4.1 onrealistisch maken is de sprong op een bepaald punt van 0 naar 1: de functies zijn discontinu.



Figuur 4.1
Itemresponsfunctie in een deterministisch model

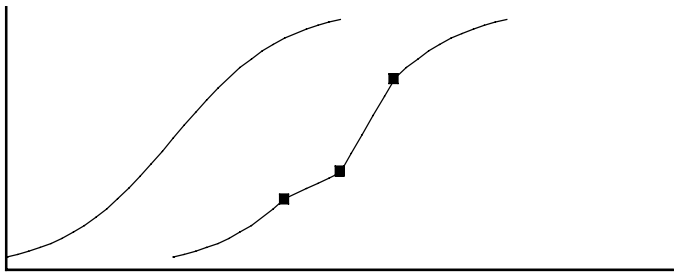
Wat we dan wel weer als een realistische eigenschap kunnen beschouwen, is dat de functies in figuur 4.1 nooit dalen: de kans op een juist antwoord wordt nooit kleiner als de vaardigheid toeneemt. We gaan deze eigenschap aanscherpen door te eisen dat de functie overal stijgend moet zijn, dat wil zeggen dat ze niet constant mag blijven in een bepaald gebied.

Samengevat stellen we de volgende eisen aan de itemresponsfunctie:

(1) $0 < f_i(\theta) < 1$;

(2) de functie is continu: de grafiek moet getekend kunnen worden zonder de pen op te tillen;

(3) de functie is strikt stijgend.



Figuur 4.2
Een 'vloeiende' en een 'hoekige' itemresponsfunctie

Figuur 4.2 toont twee grafieken die aan deze drie eisen voldoen. Een eigenschap die de twee grafieken onderscheidt is de 'hoekigheid'. Functies die dit soort hoekigheid vertonen zijn wiskundig meestal niet elegant om mee te werken. Daarom sluiten we hoekige functies uit door een vierde eis:

(4) de functie moet een vloeiend verloop hebben, of exacter uitgedrukt: de functie moet overal differentieerbaar zijn.

Hoewel de vier gestelde eisen een groot aantal functies uitsluiten, blijven er nog heel veel functies over die aan alle gestelde eisen voldoen. Door één specifieke functie te kiezen perkt men de theorie verder in tot één speciaal geval. Zo'n speciaal geval noemt men een IRT-model. Een specifieke keuze baseert men op een veelheid aan argumenten. Op deze argumenten gaan we hier niet verder in, tenzij door op te merken dat mathematische hanteerbaarheid vaak een belangrijke overweging is.

In de rest van het hoofdstuk beperken we ons tot een eenvoudig IRT-model dat in de literatuur veel aandacht heeft gekregen. Het werd in 1960 voorgesteld door de Deense statisticus G. Rasch (Rasch, 1960, 1980). Meer ingewikkelde modellen worden in hoofdstuk 5 besproken.

4.1.1 Het Raschmodel

In het Raschmodel is de itemresponsfunctie een logistische functie. De logistische functie van een argument *y* wordt gedefinieerd als

$$f(y) = \frac{\exp(y)}{1+\exp(y)}. \tag{4.4}$$

In het Raschmodel is het argument van de logistische functie het verschil $(\theta - \beta_i)$, waarbij β_i een kengetal is dat item *i* karakteriseert. Vervangen we nu in het rechterlid van (4.4) het argument *y* door dit verschil, dan krijgen we

$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1+\exp(\theta - \beta_i)}. \tag{4.5}$$

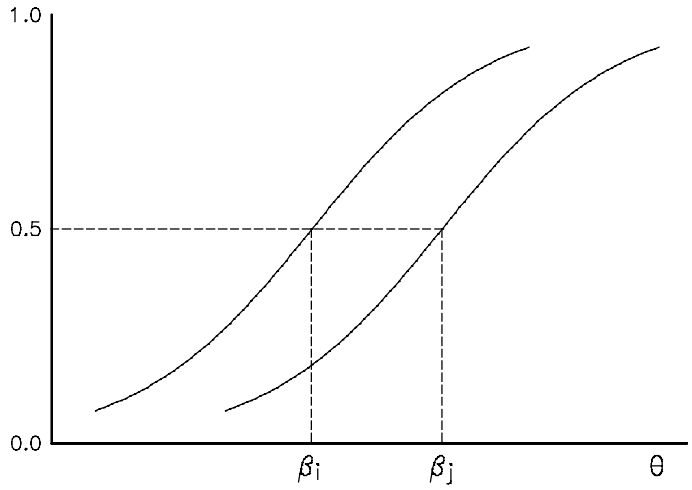
Het zal duidelijk zijn dat door de waarde van β_i te veranderen een andere functie ontstaat. Omdat we nu nog niets willen zeggen over de precieze waarde van β_i, definieert (4.5) in feite een hele familie van functies die allemaal aan de logistische functieregel voldoen. We doen een eenvoudig functieonderzoek van (4.4). Het is gemakkelijk na te gaan dat de logistische functie $f(y)$ altijd tussen 0 en 1 ligt: de teller is steeds positief en de noemer is groter dan de teller. Bovendien geldt dat $f(0) = 0.5$. Dus geldt dat

$$f_i(\beta_i) = 0.5 \tag{4.6}$$

Het is bovendien eenvoudig na te gaan dat de volgende twee limieten gelden:

$$\begin{aligned} &\lim_{\theta\to\infty} f_i(\theta) = 1, \\ &\lim_{\theta\to-\infty} f_i(\theta) = 0. \end{aligned} \tag{4.7}$$

In figuur 4.3 staan twee itemresponsfuncties afgebeeld. Twee punten van commentaar op bovenstaand functie onderzoek. Formule (4.6) betekent dat, indien de vaardigheid precies gelijk is aan het getal β_i, de kans op een juist antwoord precies 0.5 is. Omgekeerd kunnen we β_i interpreteren als de hoeveelheid vaardigheid die nodig is om een kans te hebben van 0.5 op een juist antwoord. In figuur 4.3 zien we dat meer vaardigheid vereist is om die kans te halen bij item *j* dan bij item *i*. Het is dus gerechtvaardigd om te zeggen dat β_i de moeilijkheid uitdrukt van item *i*. De parameter β_i wordt daarom vaak de moeilijkheids- parameter van het item genoemd. Omdat er in het Raschmodel met elk item slechts een parameter gemoeid is, wordt β_i ook vaak kortweg de itemparameter genoemd.



Figuur 4.3
Twee itemresponsfuncties in het Raschmodel

Het tweede commentaar heeft betrekking op (4.7). Voor zeer kleine waarden van θ is de kans bijna 0 dat een correct antwoord wordt gegeven. Dit betekent dat het Raschmodel eigenlijk ongeschikt is voor items waarvan het juiste antwoord door raden tot stand komt. Dit betekent dat extra voorzichtigheid geboden is wanneer het Raschmodel wordt toegepast bij meerkeuze-items: iemand die helemaal niets weet over het gevraagde onderwerp heeft een substantiële kans op een juist antwoord als hij gaat raden.

Een inspectie van figuur 4.3 laat zien dat de twee curven een identieke vorm hebben; ze zijn alleen verschoven ten opzichte van elkaar. Dit betekent ook dat ze elkaar nooit kruisen. Daaruit volgt dat $f_i(\theta) > f_j(\theta)$ voor elke waarde van θ. In woorden: wat ook de waarde van θ is, de kans om item *i* juist te maken is steeds groter dan de kans om item *j* juist te maken.

4.1.2 Lokale stochastische onafhankelijkheid

Formule (4.5) beschrijft het gedrag van iemand met vaardigheid θ op één item. Dit is echter niet voldoende om het Raschmodel te karakteriseren. Er moet ook nog iets gezegd worden over het gedrag, indien meer items moeten worden beantwoord. Stel dat we over vier items beschikken die precies even moeilijk zijn, en we leggen die items voor aan twee personen waarvan we weten dat ze dezelfde θ-waarde hebben. Na het beantwoorden van de eerste drie items stellen we vast dat de eerste persoon drie juiste antwoorden heeft gegeven en de tweede persoon drie onjuiste. Is het dan niet redelijk

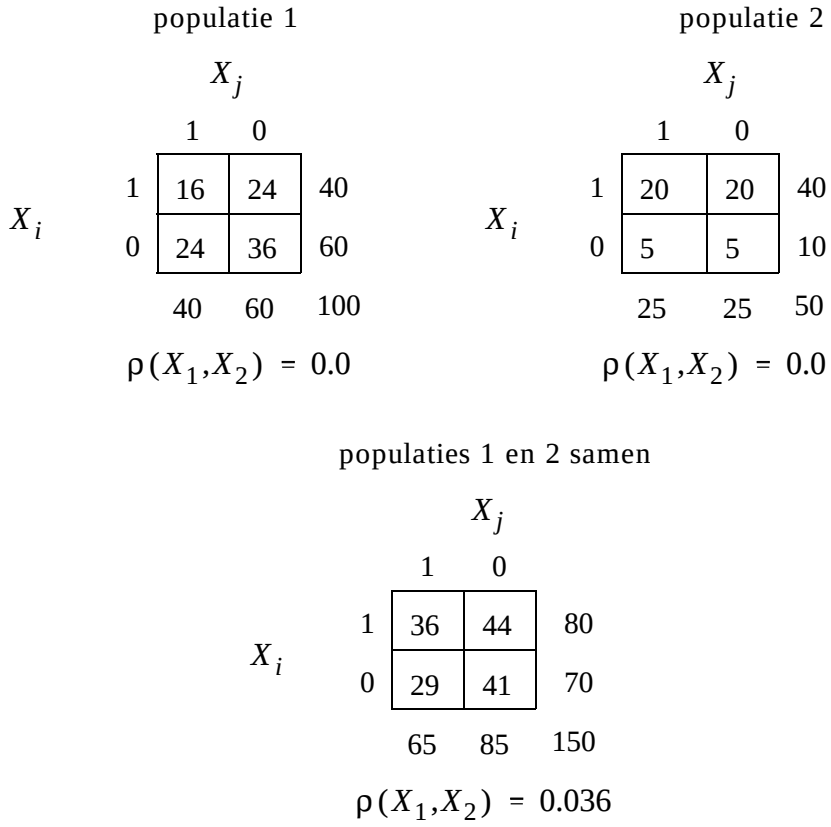
te veronderstellen dat de eerste persoon een grotere kans heeft om het vierde item juist te maken dan de tweede persoon? De eerste persoon heeft immers er blijk van gegeven vaardiger te zijn dan de tweede, gezien zijn drie juiste antwoorden. Het antwoord luidt: neen. Immers, als we aannemen dat het Raschmodel geldig is, dan hangt de kans op een juist antwoord alleen af van de vaardigheid en de moeilijkheid van het item, en in de beschreven situatie gaat het om items met dezelfde moeilijkheid en om personen met dezelfde vaardigheid. Dus moeten die kansen gelijk zijn. Kennis van antwoorden op andere items kan die kans niet veranderen. Deze redenering volgt niet automatisch uit (4.5); ze wordt toegevoegd als een onafhankelijk principe of axioma, namelijk het axioma der lokale stochastische onafhankelijkheid. Dit principe kan op verschil- lende equivalente manieren in formulevorm worden uitgedrukt. We geven twee belangrijke formules. De antwoordvariabelen X_i en X_j zijn lokaal stochastisch onafhankelijk (van elkaar) indien

$$P(X_i=1 \mid \theta \text{ en } X_j=1) = P(X_i=1 \mid \theta) = f_i(\theta), \tag{4.8}$$

of

$$P(X_i=1 \text{ en } X_j=1 \mid \theta) = P(X_i=1 \mid \theta)\, P(X_j=1 \mid \theta) = f_i(\theta)\, f_j(\theta). \tag{4.9}$$

Let wel (4.8) en (4.9) zijn niet twee verschillende voorwaarden; ze zijn equivalent en betekenen dus precies hetzelfde. De beperking 'lokaal' wijst erop dat X_i en X_j alleen onafhankelijk zijn bij gelijke θ. Daaruit volgt niet dat X_i en X_j onafhankelijk zijn van elkaar. Dus uit lokale stochastische onafhankelijkheid volgt niet dat $P(X_i=1 \text{ en } X_j=1) = P(X_i=1) \times P(X_j=1)$. Immers, indien dit waar zou zijn, dan zou de correlatie tussen de antwoorden op item *i* en item *j* nul bedragen, iets wat in het algemeen niet waar is als die items dezelfde vaardigheid meten. Het principe van de lokale stochastische onafhankelijkheid impliceert wel dat de correlatie tussen X_i en X_j nul is in alle populaties waar θ constant is. Dit geeft ons meteen een aardige manier om de correlatie tussen items te verklaren: als in een populatie de correlatie tussen item *i* en *j* niet nul is, dan komt dat doordat de vaardigheid in die populatie niet constant is. Door de invloed van de vaardigheid te controleren, dat wil zeggen door de vaardigheid constant te houden verdwijnt de correlatie. We illustreren dit aan de hand van een voorbeeld. In figuur 4.4 is duidelijk te zien dat de variabelen X_i en X_j niet corre- leren in populatie 1 noch in populatie 2. Voegen we de twee populaties echter samen, dan wordt de correlatie positief.



Figuur 4.4
Een voorbeeld van lokale stochastische onafhankelijkheid

Het axioma van de lokale stochastische onafhankelijkheid is zeer belangrijk in de IRT, maar het is erg moeilijk om te controleren of eraan voldaan is. We kunnen namelijk niet te werk gaan op de manier zoals weergegeven in figuur 4.4. Dit zou vereisen dat we de totale steekproef zouden kunnen opdelen in groepjes personen die dezelfde θ-waarde hebben. Doch θ kennen we niet, dus is deze benadering onmogelijk. Voor de toetsconstructeur is het belang- rijk het axioma niet te schenden door items te maken die functioneel afhankelijk zijn van elkaar, waar een juist antwoord op een item een juist antwoord op een ander item veronder- stelt.

4.2 Het schatten van de parameters in het Raschmodel

4.2.1 Grootste-aannemelijkheidsschatters: een voorbeeld

Door het Raschmodel als model voor het beantwoorden van de items aan te nemen zijn we natuurlijk nog niet klaar met het werk. Om (4.4) uit te rekenen moeten we een getalswaarde invullen voor θ en voor β_i en die getallen kennen we niet. θ en β_i worden parameters ge- noemd en men gebruikt de observaties om schattingen te maken van de parameters.

Er zijn verschillende manieren om parameters te schatten. Hier wordt er één besproken, namelijk de grootste-aannemelijkheidsmethode. In het Engels: maximum likelihood, afgekort als ML. De ML-methode wordt verreweg het meest gebruikt in de IRT-literatuur; ze heeft bepaalde theoretische voordelen waarop later uitvoerig wordt teruggekomen. We leggen de methode uit aan de hand van een voorbeeld. Een onzuiver muntstuk wordt vijf maal opgegooid, waarbij de uitkomst munt als een succes beschouwd wordt en de uitkomst kruis als een mislukking. We definiëren weer toevalsvariabelen X_i als

$$X_i = \begin{cases} 1 \text{ indien munt bij de } i\text{-de beurt,} \\ \\ 0 \text{ indien kruis bij de } i\text{-de beurt, } (i = 1, \ldots, 5). \end{cases}$$

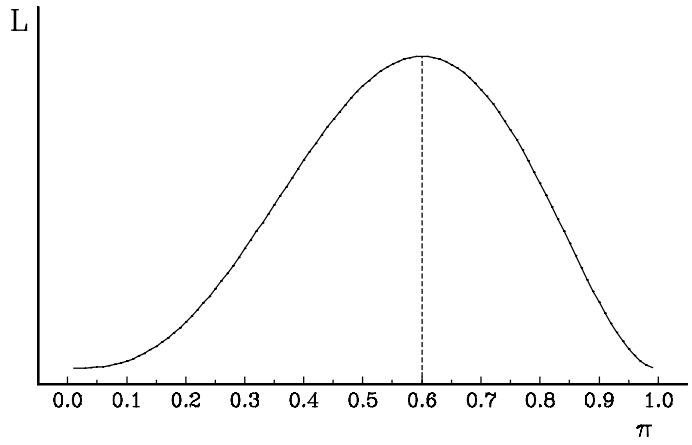
Het model is zeer simpel. Het zegt dat de kans op succes bij opgooien gelijk is aan π, waarbij π een getal is tussen 0 en 1. Wij willen de uitkomst van ons kleine experimentje gebruiken om π te schatten. Stel dat we de volgende uitkomst waarnemen: (1 0 1 1 0). De kans op die uitkomst is

$$\begin{aligned} P(X_1=1, X_2=0, X_3=1, X_4=1, X_5=0; \pi) &= \pi(1-\pi)\pi\pi(1-\pi) \\ &= \pi^3(1-\pi)^2. \end{aligned} \tag{4.10}$$

Formule (4.10) kunnen we op twee manieren bekijken. We kunnen de uitkomst van het experiment als argument van de functie P bekijken en voor alle mogelijke uitkomsten van het experiment een uitdrukking vinden die analoog is aan het rechterlid van (4.10). Dan vinden we een aantal uitdrukkingen waarin π verschijnt als een vast, hoewel nog onbekend, getal. Daarom staat π na de ';' in het linkerlid van (4.10). We kunnen (4.10) echter ook bekijken als een functie van π, waarbij we de uitkomst van ons experiment beschouwen als een gegeven. Voor elke waarde van π die we dan invullen, krijgen we als uitkomst hoe waar- schijnlijk onze observaties zijn, als π die waarde aanneemt. De functie (4.10) zo bekeken noemt men de aannemelijkheidsfunctie (Engels: likelihood function) en die wordt gegeven door

$$L(\pi; (1\,0\,1\,1\,0)) = P((1\,0\,1\,1\,0); \pi). \tag{4.11}$$

De grafiek van het rechterlid van (4.11) is weergegeven in figuur 4.5.



Figuur 4.5
Aannemelijkheidsfunctie voor de observatie (1 0 1 1 0)

De ML-schatting van π is die waarde van π waarvoor de aannemelijkheidsfunctie zo groot mogelijk wordt, dat wil zeggen die waarde waarvoor de gegeven observaties de grootste waarschijnlijkheid hebben. In het voorbeeld is dit 0.6 zoals makkelijk uit figuur 4.5 kan worden afgelezen. Natuurlijk zal men niet steeds een grafiek van de aannemelijkheidsfunctie maken om de schatting te bepalen. Men gebruikt een standaardtechniek, die hier even kort wordt besproken.

Aan de manier waarop (4.10) is opgesteld kan men duidelijk zien dat de volgorde waarin successen en mislukkingen zich voordoen tijdens het experiment niet belangrijk is voor de aannemelijkheidsfunctie; alleen het aantal successen en mislukkingen telt. Indien er *n* keer wordt opgegooid en er zijn *s* successen, dan zijn er *n-s* mislukkingen. Stellen we de uitkomsten van een experiment voor door $\boldsymbol{x} = (x_1,\ldots,x_n)$ dan krijgen we als algemene uitdrukking voor de aannemelijkheidsfunctie

$$L(\pi;\boldsymbol{x}) = \pi^s(1-\pi)^{n-s}, \tag{4.12}$$

waarin $s = \sum_{i=1}^{n} x_i$. Om het maximum van (4.12) te zoeken kiest men gewoonlijk een andere
functie waarvan men weet dat ze monotoon is met de aannemelijkheidsfunctie. De functie die meestal wordt gebruikt is de logaritme van de aannemelijkheidsfunctie:

$$\ln L(\pi;\boldsymbol{x}) = s\ln\pi + (n-s)\ln(1-\pi). \tag{4.13}$$

Een standaardmanier om een maximum van een functie te zoeken is, de eerste afgeleide van die functie te bepalen, die afgeleide gelijk te stellen aan nul en de aldus ontstane vergelijking op te lossen naar de onbekende parameter. Deze vergelijking wordt schattingsvergelijking of aannemelijkheidsvergelijking genoemd. De eerste afgeleide van (4.13) is

$$\frac{d \ln L(\pi;\boldsymbol{x})}{d\pi} = \frac{s}{\pi} - \frac{n-s}{1-\pi}. \tag{4.14}$$

Gelijkstellen van (4.14) aan 0 geeft als oplossing

$$\hat{\pi} = \frac{s}{n}. \tag{4.15}$$

Het rechterlid van (4.15) is een functie van de gegevens. We zien dus dat we een algemene oplossing krijgen voor het muntexperiment: de grootste-aannemelijkheids-schatter is het aantal successen gedeeld door het aantal keren opgooien. De functie s/n wordt de schatter genoemd. De waarde die die functie aanneemt in een concreet geval wordt de schatting genoemd. In het voorbeeld is de schatting van π dus gelijk aan 0.6. Het dakje boven het parametersymbool wordt gebruikt om aan te geven dat het hier niet gaat om de echte waarde van π, maar om een schatter of schatting. De schatter is een functie van het aantal successen, en dit aantal is een toevalsvariabele; dus is de schatter ook een toevalsvariabele, en de schatting zelf zal van experiment tot experiment verschillen.

Omdat we meestal niet een zeer groot aantal experimenten uitvoeren maar slechts één, blijven we met de vraag zitten of de schatting die we in een concreet geval voor π krijgen wel een goede schatting is. Bovendien is er nog een ander probleem: de oplossing (4.15) garandeert ons alleen dat de eerste afgeleide van (4.14) 0 is indien $\pi = s/n$, doch daaruit volgt niet automatisch dat dit punt met een maximum overeenkomt. Daartoe moeten we hogere afgeleiden van (4.14) onderzoeken. Indien de tweede afgeleide negatief is op het punt waar de eerste afgeleide nul wordt weten we dat we te doen hebben met een maximum. De tweede afgeleide van de log-aannemelijkheidsfunctie is gegeven door

$$\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2} = -\frac{s}{\pi^2} - \frac{n-s}{(1-\pi)^2}, \tag{4.16}$$

en deze functie is negatief voor alle waarden van π in het interval (0,1). (De gevallen waar $\pi = 0$ en $\pi = 1$ laten we buiten beschouwing.) De oplossing (4.15) komt dus overeen met een maximum van de aannemelijkheidsfunctie.

De tweede afgeleide kunnen we ook gebruiken om iets te zeggen over de nauwkeurigheid van de ML-schatter van π. In de theoretische statistiek zijn belangrijke resultaten bekend over de statistische eigenschappen van ML-schatters. Hoewel deze resultaten niet altijd geldig zijn, zijn ze wel bruikbaar voor de modellen die in dit boek worden behandeld. Bovendien staan deze resultaten bekend als 'asymptotische' resultaten, dit wil zeggen dat ze strikt genomen alleen geldig zijn als $n \to \infty$. In de praktijk kunnen ze echter goed gebruikt worden als de steekproef niet al te klein is. Het belangrijkste resultaat luidt:

> De ML-schatter is asymptotisch normaal verdeeld met gemiddelde de werkelijke parameter π van het model en als variantie één gedeeld door de informatiefunctie. (Zie bijvoorbeeld Kendall & Stuart, 1973.)

De informatiefunctie $I(\pi)$ met betrekking tot de parameter π is gedefinieerd als

$$I(\pi) = -\mathscr{E}\left[\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2}\right], \tag{4.17}$$

waarbij de verwachte waarde genomen dient te worden over alle mogelijke steekproeven (met vaste n). In het voorbeeld met het muntstuk geeft dit

$$\begin{aligned} I(\pi) &= -\mathscr{E}\left[\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2}\right] \\ &= \frac{\mathscr{E}(s)}{\pi^2} + \frac{n - \mathscr{E}(s)}{(1-\pi)^2} \\ &= \frac{n\pi}{\pi^2} + \frac{n(1-\pi)}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)}. \end{aligned} \tag{4.18}$$

Uit (4.18) en het bovengenoemde resultaat volgt onmiddellijk dat de schatter $\hat{\pi} = s/n$ asymptotisch normaal verdeeld is met gemiddelde π en variantie $\pi(1-\pi)/n$, een resultaat dat in elke cursus statistiek gepresenteerd wordt. Om de variantie uit te rekenen moeten we echter de waarde van π kennen. Omdat die niet bekend is, vult men daarvoor de ML-schatting in van π. Dit geeft dus als resultaat

$$\sigma^2(\hat{\pi}) \approx \frac{1}{I(\hat{\pi})} = \frac{\hat{\pi}(1-\hat{\pi})}{n}. \tag{4.19}$$

Het teken '$\approx$' geeft aan dat de gelijkheid slechts asymptotisch geldt; de echte standaardfout bij een eindige steekproef is in de regel groter dan door (4.19) is

aangegeven. De standaardfout (verder afgekort als *SE*, van het Engelse standard error), dit is de vierkantswortel uit (4.19), kan gebruikt worden om bijvoorbeeld betrouwbaarheidsintervallen voor de parameter te berekenen. Passen we (4.19) toe op het voorbeeld, dan vinden we $\sigma^2(\hat{\pi}) \approx .24/5 = .048$. Het 95%-betrouwbaarheidsinterval is dus gegeven door $(\hat{\pi} - 1.96 \times \sqrt{0.48}\,, \hat{\pi} + 1.96 \times \sqrt{0.48}) = (0.17,1.03)$. Dit grote betrouwbaarheidsinterval, dat zich hier uitstrekt buiten het toegestane bereik van de parameter, is te wijten aan de uiterst kleine steekproef, die ons niet veel informatie over de parameter oplevert. Hadden we 50 keer opgegooid met het muntstuk, dan hadden we bij 30 successen een variantie gekregen van .0048, en een standaardfout die $10^{½} = 3.16$ zo klein was, en dus ook een betrouwbaar- heidsinterval dat 3.16 kleiner is: (0.46,0.74).

In de literatuur wordt nog een andere manier gebruikt om een schatting van de standaardfout te verkrijgen. In plaats van de verwachte waarde te nemen van minus de tweede afgeleide van de log-aannemelijkheidsfunctie, neemt men gewoon minus de tweede afgeleide van de log-aannemelijkheidsfunctie zelf. Deze functie, geëvalueerd op de ML-schatting, wordt de geobserveerde-informatiefunctie genoemd. Het symbool dat hiervoor gebruikt wordt is *J*. Uit (4.15) volgt dat $s = n\hat{\pi}$. Dus krijgen we, door invullen in (4.16)

$$J(\hat{\pi}) = \frac{n\hat{\pi}}{\hat{\pi}^2} + \frac{n - n\hat{\pi}}{(1-\hat{\pi})^2} = \frac{n}{\hat{\pi}(1-\hat{\pi})}. \qquad (4.20)$$

Het feit dat we voor de informatiefunctie, geëvalueerd op de ML-schatter, en voor de geobserveerde informatiefunctie hetzelfde resultaat krijgen is niet toevallig en heeft te maken met een speciale eigenschap van de log-aannemelijkheidsfunctie. Het is niet moeilijk na te gaan dat de log-aannemelijkheidsfunctie geschreven kan worden als

$$\ln L(\pi; \boldsymbol{x}) = s \ln \frac{\pi}{1-\pi} + n \ln (1-\pi). \qquad (4.21)$$

De eerste term in het rechterlid van (4.21) is een produkt van twee factoren: de eerste factor is een functie van de gegevens (*s*) en de tweede factor is een functie van de parameter. De
tweede term is alleen een functie van de parameter π (*n* dient beschouwd te worden als een constante). Dit is een iets gespecialiseerde vorm van een meer algemene vorm van de log-aannemelijkheidsfunctie. Indien men een model beschouwt met meer dan één parameter, bijvoorbeeld *k*, waarbij de parameters verzameld zijn in de *k*-vector π, en men kan de log-aannemelijkheidsfunctie schrijven als

$$\ln L(\pi;\boldsymbol{x}) = \sum_{i=1}^{k} A_i(\boldsymbol{x})\, B_i(\pi) + C(\pi) + D(\boldsymbol{x}), \tag{4.22}$$

waarin A_i en D functies zijn van de gegevens maar niet van de parameters, en B_i en C functies zijn van de parameters maar niet van de gegevens, dan zegt men dat de log-aannemelijkheidsfunctie (of het model) behoort tot de exponentiële familie. Formule (4.21) is gemakkelijk te herkennen als een speciaal geval van (4.22), met $k = 1$, $A_1 = s$, $B_1 = \ln[\pi/(1-\pi)]$, $C = n\ln(1-\pi)$ en $D = 0$. De exponentiële familie heeft veel prettige eigenschappen, en één ervan is dat de informatiefunctie, geëvalueerd op de ML-schatter, en de geobserveerde informatiefunctie gelijk zijn aan elkaar.

Tenslotte nog een opmerking over de functies A_i in (4.22). Deze functies worden de minimaal voldoende steekproefgrootheden, in het Engels: minimal sufficient statistics, genoemd voor de functies $B_i(\pi)$. Dat een steekproefgrootheid voldoende is om de parameter te schatten, betekent dat we van de observaties niet méér gebruiken dan door deze grootheid wordt aangegeven. Bij het muntstuk experiment is het aantal successen voldoende om de parameter π te schatten; de precieze afwisseling van successen en mislukkingen levert geen bijkomende informatie over de parameter. Op de term 'minimaal' dienen we echter nog even in te gaan. Stel dat de k-de functie $B_k(\pi)$ in (4.22) kan geschreven worden als een lineaire combinatie van de $k-1$ andere functies $B_i(\pi)$, dat wil zeggen dat er getallen $\alpha_1,\ldots,\alpha_{k-1}$ bestaan zodat

$$\begin{aligned} B_k(\pi) &= \alpha_1 B_1(\pi) + \ldots + \alpha_{k-1} B_{k-1}(\pi) \\ &= \sum_{i=1}^{k-1} \alpha_i B_i(\pi), \end{aligned} \tag{4.23}$$

dan kan (4.22) geschreven worden als

$$\begin{aligned} \ln L(\pi;\boldsymbol{x}) &= \sum_{i=1}^{k-1} A_i(\boldsymbol{x})\, B_i(\pi) + A_k(\boldsymbol{x}) \sum_{i=1}^{k-1} \alpha_i B_i(\pi) + C(\pi) + D(\boldsymbol{x}) \\ &= \sum_{i=1}^{k-1} [A_i(\boldsymbol{x}) + \alpha_i A_k(\boldsymbol{x})]\, B_i(\pi) + C(\pi) + D(\boldsymbol{x}). \end{aligned} \tag{4.24}$$

Doch de factor tussen [ ] in het rechterlid van (4.24) is geen functie van de parameters, en dus is (4.24) een log-aannemelijkheidsfunctie uit de exponentiële familie, maar nu met $k-1$ parameters. Op analoge manier kan men soms het aantal parameters verminderen door aan te tonen dat een functie $A_i(\boldsymbol{x})$ lineair afhankelijk is van de

andere *A*-functies. Als we spreken over het aantal parameters in een model, dan zullen we altijd het aantal bedoelen waarvoor een verdere restrictie als gegeven in (4.23) niet meer mogelijk is. Deze parameters worden ook wel aangeduid als vrije parameters.

4.2.2 JML-schatting in het Raschmodel

In het Raschmodel kunnen we proberen op een soortgelijke manier te werk te gaan als in de vorige paragraaf. De principes blijven dezelfde, er is alleen een complicatie omdat we nu niet één parameter moeten schatten, maar verschillende tegelijkertijd. Nemen we een toets bestaande uit *k* items af aan *n* personen, dan moeten we *n* θ-parameters schatten en *k* itemparameters. De J in JML staat voor 'joint'. Men gebruikt deze aanduiding niet om aan te geven dat er meer parameters geschat moeten worden, maar om aan te geven dat de twee soorten parameters, persoonsparameters en itemparameters, tegelijkertijd geschat worden. Om de aannemelijkheidsfunctie op te stellen moeten we de notatie iets uitbreiden. De toevalsvariabele X_{vi} verwijst naar het antwoord van persoon *v* op item *i*. De waarden die die toevalsvariabele kan aannemen, 0 of 1, zullen we in het algemeen aanduiden met $\boldsymbol{x}_{vi}$. Willen we verwijzen naar de antwoorden van persoon *v*, dan wordt dit aangeduid met $\boldsymbol{x}_v$, en willen we verwijzen naar alle antwoorden van alle personen in de steekproef dan wordt dit aangeduid met $\boldsymbol{X}$.

Beschouw eerst als voorbeeld een steekproef van een persoon *v*, met $\theta = \theta_v$, en een toets van k=3 items. Veronderstel dat we de antwoorden (1,0,1) hebben geobserveerd. Gebruik makend van het principe van de lokale stochastische onafhankelijkheid en van formule (4.3), kan de aannemelijkheidsfunctie voor dit antwoordpatroon geschreven worden als

$$L(\beta_1,\beta_2,\beta_3,\theta_v;(1\,0\,1)) = f_1(\theta_v)(1-f_2(\theta_v))f_3(\theta_v). \qquad (4.25)$$

Merk op dat bovenstaand produkt bestaat uit k=3 factoren, dat met een juist antwoord op item *i* een factor $f_i(\theta_v)$ overeenkomt, en met een verkeerd antwoord een factor $(1-f_i(\theta_v))$. Om een algemene formule te verkrijgen, wordt het produkt in (4.25) uitgebreid tot 2*k* factoren, twee per item. Het produkt van die twee factoren heeft de gedaante

$$[f_i(\theta_v)]^{x_{vi}}[1-f_i(\theta_v)]^{1-x_{vi}}.$$

Indien $x_{vi} = 1$ is dit produkt gelijk aan $f_i(\theta_v)$, en indien $x_{vi} = 0$, is het produkt gelijk aan $(1 - f_i(\theta_v))$. Duiden we nu met β de vector $(\beta_1,...,\beta_k)$ aan, dan krijgen we als directe veralgemening van (4.25):

$$L(\beta,\theta_v;\boldsymbol{x}_v) = \prod_{i=1}^{k} [f_i(\theta_v)]^{x_{vi}} [1-f_i(\theta_v)]^{1-x_{vi}}. \tag{4.26}$$

Veralgemenen we dit nu tot een steekproef van *n* personen. Elke persoon levert een aannemelijkheidsfunctie op van de gedaante (4.26). De aannemelijkheidsfunctie voor alle gegevens samen is het produkt van de aannemelijkheidsfunctie voor alle antwoordpatronen afzonderlijk. Dit is waar indien de antwoorden van de personen onafhankelijk zijn van elkaar. Let wel, de reden is niet de lokale stochastische onafhankelijkheid, want we kunnen er niet van uitgaan dat alle personen de zelfde θ-waarde hebben. Onafhankelijkheid betekent hier dat de antwoorden van de ene persoon geen informatie bevatten over de antwoorden van een andere persoon. Dit soort onafhankelijkheid wordt in de testtheorie experimentele onafhankelijkheid genoemd. Duiden we de vector $(\theta_1,...,\theta_n)$ aan met θ, dan vinden we

$$L(\beta,\theta;\boldsymbol{X}) = \prod_{v=1}^{n}\prod_{i=1}^{k} [f_i(\theta_v)]^{x_{vi}} [1-f_i(\theta_v)]^{1-x_{vi}}. \tag{4.27}$$

Substitueren we nu (4.5) in (4.27), en nemen we de logaritme, dan vinden we

$$\ln L(\beta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v\theta_v + \sum_{i=1}^{k} t_i(-\beta_i) - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln[1 + \exp(\theta_v - \beta_i)], \tag{4.28}$$

waarin

$$s_v = \sum_{i=1}^{k} x_{vi}, \qquad t_i = \sum_{v=1}^{n} x_{vi}.$$

Het is makkelijk in te zien dat (4.28) een log-aannemelijkheidsfunctie uit de exponentiële familie is, met s_v, $v = 1,...,n$ en t_i, $i = 1,...,k$, de voldoende steekproef-grootheden voor respectievelijk θ_v, $v = 1,...,n$, en $(-\beta_i)$, $i = 1,...,k$. De laatste term in (4.28) komt overeen met de functie *C* in (4.22). Er geldt echter:

$$\sum_v s_v = \sum_i t_i,$$

dat wil zeggen dat er een lineaire restrictie op de grootheden s_v en t_i ligt. Er zijn dus niet $k + n$ maar hoogstens $k + n - 1$ vrije parameters; meer parameters kunnen dus ook

niet ge- schat worden. Dit betekent dat het Raschmodel in zijn algemeenheid niet schatbaar is, of zoals men het ook uitdrukt: het model is niet geïdentificeerd. Dit valt reeds af te leiden uit de itemresponsfunctie (4.5). Stel dat we van alle personen θ_v en van alle items β_i kennen. Een andere, doch evenwaardige oplossing bestaat erin aan elke persoon v het getal $\theta_v^* = \theta_v + c$ en aan elk item het getal $\beta_i^* = \beta_i + c$ toe te kennen, waarbij c een willekeurige constante is. Dan geldt natuurlijk dat $\theta_v^* - \beta_i^* = \theta_v - \beta_i$, en dus blijft de itemresponsfunctie onveranderd welke waarde we ook aan c geven. Willen we zinvol over de parameters kunnen spreken dan moeten we de waarde van c vastleggen, of met ander woorden, we moeten het nulpunt van de schaal vastleggen. Dit kunnen we doen door bijvoorbeeld één van de parameters (bijvoorbeeld β_1) gelijk te stellen aan nul. Doch in dat geval zijn er nog maar $k - 1$ vrije itemparameters over, hetgeen in overeenstemming is met de bovenvermelde lineaire restrictie. Het kiezen van het nulpunt noemt men normaliseren. De meest gebruikte normalisatie is het nulpunt zo te kiezen dat $\sum_{i=1}^{k} \beta_i = 0$.

Om het maximum van (4.28) te vinden, kan men een generalisatie van de techniek toepassen die in paragraaf 4.2.1 werd besproken. Op het maximum van een functie van meerdere parameters moeten alle partiële afgeleiden gelijk zijn aan nul. De partiële afgeleide van een functie naar een parameter is de afgeleide van de functie naar die parameter, waarbij alle andere parameters als constante worden beschouwd. We hoeven deze exercitie echter niet uit te voeren omdat we gebruik kunnen maken van een resultaat dat geldig is in de exponentiële familie. Dit resultaat luidt:

> In een exponentieel familie model zijn de aannemelijkheidsvergelijkingen gegeven door de voldoende steekproefgrootheden gelijk te stellen aan hun verwachte waarde (Andersen, 1980).

Dit geeft dus voor de θ-parameters:

$$\begin{aligned} s_v &= \mathscr{E}(S_v) = \mathscr{E}\left[\sum_i X_{vi}\right] = \sum_i \mathscr{E}(X_{vi}) \\ &= \sum_i [1 \times P(X_{vi}=1 \mid \theta_v) + 0 \times P(X_{vi}=0 \mid \theta_v)] \\ &= \sum_i f_i(\theta_v), \qquad (v = 1, \ldots, n), \end{aligned} \tag{4.29}$$

waarin S_v de toevalsvariabele 'score van persoon v' aanduidt met als realisatie de geobser- veerde score s_v. Zij T_i de toevalsvariabele 'aantal juiste antwoorden gegeven op item i', dan worden de schattingsvergelijkingen voor de β-parameters gegeven door

$$t_i = \mathscr{E}(T_i) = \sum_v f_i(\theta_v), \qquad (i = 2, \ldots, k). \tag{4.30}$$

In (4.30) is geen vergelijking opgenomen voor i=1. Dit betekent dat β_1 niet beschouwd wordt als een parameter die geschat moet worden, maar als een bekende constante. De waarde die we aan β_1 geven is in principe willekeurig; wij zullen echter aannemen dat $\beta_1 = 0$. Merk op dat (4.29) en (4.30) een stelsel van vergelijkingen vormen in $k+n-1$ onbekenden. Dit stelsel kan niet expliciet worden opgelost, de oplossing wordt gezocht met een iteratieve procedure, waarbij in elke iteratie aan de parameters waarden worden toegekend die de oplossing steeds dichter benaderen. Op de technische aspecten van deze oplossingsmethode gaan we hier niet in.

Er zijn echter twee problemen verbonden met het stelsel gevormd door (4.29) en (4.30). Het eerste is gemakkelijk duidelijk te maken. Stel dat er een persoon v is in de steekproef die alle items juist heeft beantwoord. Dan geldt dat het linkerlid in (4.29) gelijk is aan k. Het rechterlid bestaat uit k termen die alle strikt kleiner zijn dan 1, dus hun som is kleiner dan k, welke waarden men ook voor de parameters invult. Een analoog probleem krijgt men wanneer $s_v = 0$. Bij de vergelijkingen (4.30) geldt hetzelfde argument indien $t_i = n$ of $t_i = 0$. In deze gevallen bestaat er dus geen schatter van de parameter.

Het tweede probleem is van theoretische aard en heeft betrekking op een eigenschap van schatters die men consistentie noemt (Kendall & Stuart, 1973). Ruwweg betekent consistentie dat, hoe meer informatie men verzamelt over een parameter door de steekproef steeds groter te maken, des te nauwkeuriger de schatting moet zijn en in de limiet, bij $n \to \infty$ is de kans dat men de parameter juist schat gelijk aan 1. In het geval van het Raschmodel treedt er echter een complicatie op: om meer informatie te verzamelen over itemparameters dient men de toets steeds bij nieuwe personen af te nemen, doch elke persoon die men aan de steekproef toevoegt brengt zijn eigen onbekende θ-parameter mee. Dit wil zeggen dat de omvang van het probleem, het aantal te schatten parameters, even snel groeit als de steekproef. Het gevolg hiervan is dat de JML-schatters van de itemparameters niet consistent zijn. Bovendien gelden de asymptotische resultaten over de standaardfout, die in paragraaf 4.2.1. werden besproken, hier niet automatisch. Dit maakt de JML-schattingsmethode oninteressant. Als men echt in de itemparameters is geïnteresseerd, dan is het veel handiger naar een schattingsmethode te zoeken waarbij men geen last meer heeft van het steeds groeiende aantal θ-parameters. Deze parameters, waar men in eerste instantie niet zo in geïnteresseerd is, maar die toch in het model aanwezig zijn worden in de literatuur aangeduid met de term ’nuisance parameters’. De andere parameters waarin men wel is geïnteresseerd worden structurele parameters genoemd.

In de literatuur zijn verschillende methodes bekend om de 'nuisance parameters' kwijt te raken. In de twee volgende subparagrafen worden twee van deze methodes besproken.

4.2.3 CML-schatting in het Raschmodel

Het is nuttig om even het volgende gedachtenexperiment uit te voeren. De itemresponsfunctie is een conditionele kans om een juist antwoord te geven op een item. Stel nu dat we er in zouden slagen een grote steekproef samen te stellen van personen die allemaal dezelfde θ-waarde hebben, zeg θ_m. Indien aan al die personen hetzelfde item *i* zou worden voorgelegd, dan zal een proportie $p_i(\theta_m)$ het item juist beantwoorden. Deze proportie is een schatting van de conditionele kans $f_i(\theta_m)$ en uit (4.5) volgt dat, als we deze schatter invullen en de logaritme nemen,

$$\hat{\beta}_i = \theta_m - \ln \frac{p_i(\theta_m)}{1 - p_i(\theta_m)}.$$

Passen we deze methode toe op twee items, *i* en *j*, bij dezelfde steekproef, dan volgt uit het bovenstaande direct dat

$$\hat{\beta}_i - \hat{\beta}_j = \ln \frac{p_j(\theta_m)[1 - p_i(\theta_m)]}{p_i(\theta_m)[1 - p_j(\theta_m)]}. \tag{4.31}$$

Dit wil zeggen dat we een schatting krijgen van het verschil tussen twee itemparameters die onafhankelijk is van de θ-parameter, want de proportie $p_i(\theta_m)$ is een direct geobserveerde grootheid. Het bezwaar tegen deze methode is echter dat ze principieel niet uitgevoerd kan worden, omdat de θ-waarde van een persoon niet observeerbaar is; dat wil zeggen dat we geen groep van personen met allemaal dezelfde θ kunnen vormen. Wat echter wel uitvoerbaar is, is het indelen in groepen van personen met dezelfde toetsscore *s*. We bekijken eerst een voorbeeld.

Veronderstel dat $k = 3$ en beschouw het antwoordpatroon (1 0 1). De score *s* van dit ant- woordpatroon is 2. Nu zijn er exact drie mogelijke antwoordpatronen met score 2, namelijk (1 0 1), (1 1 0) en (0 1 1). Conditioneren op score 2 betekent dat we reeds weten dat een van die drie antwoordpatronen is opgetreden, en nu willen we weten wat

de kans is dat (1 0 1) is opgetreden, als alleen die drie mogelijk zijn. De formule hiervoor is

$$P(1\,0\,1|s{=}2,\theta) = \frac{P(1\,0\,1|\theta)}{P(1\,0\,1|\theta) + P(1\,1\,0|\theta) + P(0\,1\,1|\theta)}. \tag{4.32}$$

Bekijken we nu even twee equivalente formules voor het Raschmodel:

$$P(X_i{=}1|\theta) = f_i(\theta) = \frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)}, \tag{4.33}$$

en

$$P(X_i{=}0|\theta) = 1 - f_i(\theta) = \frac{1}{1+\exp(\theta-\beta_i)}. \tag{4.34}$$

Als we de aannemelijkheidsfunctie opstellen moeten we produkten nemen van uitdrukkingen met de gedaante (4.33) voor juiste antwoorden of (4.34) voor foute antwoorden. Merk op dat de noemers van (4.33) en (4.34) identiek zijn. De noemer van het produkt is dus onafhankelijk van het specifieke antwoordpatroon. Stel deze noemer voor door het symbool *K*. Beschouw nu de kans op het antwoordpatroon (1 0 1):

$$P(1\,0\,1|\theta) = \frac{\exp(\theta)\exp(-\beta_1)\exp(\theta)\exp(-\beta_3)}{K} = \frac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}. \tag{4.35}$$

In de teller van (4.35) komt 2θ voor in de exponent. Het is duidelijk dat die 2 daar staat, omdat het over een antwoordpatroon gaat met precies 2 juiste antwoorden. Doch dit is ook het geval voor de antwoordpatronen (1 1 0) en (0 1 1). Dan is het niet moeilijk in te zien dat

$$\begin{aligned} &P(1\,0\,1\,|s=2,\theta) \\ &= \frac{\dfrac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}}{\dfrac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K} + \dfrac{\exp(2\theta)\exp(-\beta_1-\beta_2)}{K} + \dfrac{\exp(2\theta)\exp(-\beta_2-\beta_3)}{K}} \\ &= \frac{\exp(-\beta_1-\beta_3)}{\exp(-\beta_1-\beta_3)+\exp(-\beta_1-\beta_2)+\exp(-\beta_2-\beta_3)}. \end{aligned} \tag{4.36}$$

Het belangrijke aspect van (4.36) is dat het rechterlid onafhankelijk is van θ en alleen nog een functie van de itemparameters. Bij de vereenvoudiging van (4.36), dat wil zeggen de overgang van het tweede lid naar het derde lid, merken we dat niet alleen de noemers K verdwijnen, maar ook de uitdrukking 2θ. Dit kon alleen maar door ervoor te zorgen dat θ telkens met hetzelfde getal, 2, werd vermenigvuldigd. Maar 2 is precies de score die met de drie beschouwde antwoordpatronen is geassocieerd. De 'truc' om θ te laten verdwijnen werkt dus alleen maar als we conditioneren op de score.

De uitdrukking (4.36), maar nu beschouwd als een functie van de β-parameters, noemen we de conditionele aannemelijkheidsfunctie voor het patroon (1 0 1). Om een algemene formule op te stellen voor de conditionele aannemelijkheid is het handig over te gaan op een andere parametrisering. Definieer

$$\varepsilon_i = \exp(-\beta_i), \quad (i=1,\ldots,k). \tag{4.37}$$

Met deze parameters kan (4.36) geschreven worden als

$$P(1\,0\,1\,|s=2,\theta) = \frac{\varepsilon_1\varepsilon_3}{\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3} = \frac{\prod_{i=1}^{k}\varepsilon_i^{x_i}}{\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3}. \tag{4.38}$$

De noemer in het rechterlid van (4.38) heeft een merkwaardige structuur: het is een som van drie termen, en elke term is een produkt van twee parameters. De indices van de parameters in elke term kan men opvatten als een aanduiding van de items die men juist moet hebben om een score van 2 te behalen. Er zijn drie termen omdat men slechts op drie verschillende manieren een score van 2 kan behalen. In het algemeen, bij k items en een score s $(s=0,1,\ldots,k)$, zijn er $(k!)/[s!(k-s)!]$ manieren om een score s te behalen. De noemer in de overeenkomstige formule voor de conditionele aannemelijkheid zal dus uit even zo veel termen bestaan, en elke term bestaat uit een produkt van s ε-parameters, waarvan de indices aangeven welke items juist werden beantwoord om de score s te behalen. De noemer is dus een functie van de ε-parameters, en deze functie draagt de naam 'symmetrische basisfunctie'. Voor elke score is er een andere functie; de aanduiding van de score wordt de 'orde'van de functie genoemd. Definiëren we $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_k)$, dan worden de symmetrische basisfuncties van de orde s aangeduid als $\gamma_s(\boldsymbol{\varepsilon})$. Hun definitie is

$$
\begin{aligned}
\gamma_0(\boldsymbol{\varepsilon}) &= 1,\\
\gamma_1(\boldsymbol{\varepsilon}) &= \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_k,\\
\gamma_2(\boldsymbol{\varepsilon}) &= \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \ldots + \varepsilon_1\varepsilon_k + \varepsilon_2\varepsilon_3 + \ldots + \varepsilon_{k-1}\varepsilon_k,\\
&\vdots\\
\gamma_k(\boldsymbol{\varepsilon}) &= \varepsilon_1\varepsilon_2\ldots\varepsilon_k,
\end{aligned}
\tag{4.39}
$$

$\gamma_s(\boldsymbol{\varepsilon}) = 0$ indien $s < 0$ *of* $s > k$.

De conditionele aannemelijkheidsfunctie, gegeven dat de score gelijk is aan s kunnen we nu dus algemeen schrijven als

$$
L(\boldsymbol{\varepsilon}; \boldsymbol{x} \mid s) = \frac{\prod_{i=1}^{k} \varepsilon_i^{x_i}}{\gamma_s(\boldsymbol{\varepsilon})}. \tag{4.40}
$$

De conditionele aannemelijkheidsfunctie voor alle geobserveerde antwoordpatronen samen is het produkt van soortgelijke uitdrukkingen:

$$
L(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s}) = \frac{\prod_{v=1}^{n} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\prod_{v=1}^{n} \gamma_{s_v}(\boldsymbol{\varepsilon})}, \tag{4.41}
$$

waarin $\boldsymbol{s} = (s_1, \ldots, s_n)$.

Om de schattingsvergelijkingen op te stellen, hebben we de partiële afgeleiden nodig van de γ-functies naar de ε-parameters. Neem als voorbeeld

$$
\gamma_3(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4) = \varepsilon_1\varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_4 + \varepsilon_1\varepsilon_3\varepsilon_4 + \varepsilon_2\varepsilon_3\varepsilon_4
$$

en beschouw de partiële afgeleide naar ε_2. Van de term in de uitdrukking hierboven die ε_2 niet bevat is de partiële afgeleide nul, en van de termen die ε_2 wel bevatten is de partiële afgeleide het produkt van de andere ε-parameters. Dus

$$
\frac{\partial \gamma_3(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)}{\partial \varepsilon_2} = \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4 + \varepsilon_3\varepsilon_4,
$$

doch dit is eveneens een symmetrische basisfunctie, maar nu van orde 2 en van de parameters (ε_1, ε_3, ε_4). De parameter waarnaar gedifferentieerd wordt, is uit het rijtje weggevallen. In het algemeen krijgen we dan ook de uitdrukking

$$
\frac{\partial \gamma_s(\boldsymbol{\varepsilon})}{\partial \varepsilon_i} = \gamma_{s-1}^{(i)}(\boldsymbol{\varepsilon}), \tag{4.42}
$$

waarbij de *(i)* in superscript aanduidt dat ε_i niet meer tot het argument van de γ-functie behoort.

De logaritme van (4.41) is

$$\ln L(\boldsymbol{\varepsilon}; \boldsymbol{x} \mid \boldsymbol{s}) = \sum_i t_i \ln \varepsilon_i - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}), \tag{4.43}$$

waarin weer duidelijk de structuur van de exponentiële familie tot uiting komt: de grootheden t_i zijn de voldoende steekproefgrootheden voor de parameters $\ln(\varepsilon_i)$. Dus ook de conditionele verdeling van X gegeven $\boldsymbol{s}$ behoort tot deze familie. Stellen we de partiële afgeleiden van (4.43) naar ε_i gelijk aan 0, dan krijgen we als schattingsvergelijkingen

$$t_i = \sum_v \frac{\varepsilon_i \gamma^{(i)}_{s_v - 1}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}, \qquad (i = 2, \ldots, k). \tag{4.44}$$

Gebruik makend van een reeds eerder vermelde eigenschap van de exponentiële familie, kunnen we echter ook schrijven dat

$$t_i = \mathscr{E}(T_i \mid \boldsymbol{s}) = \sum_v \pi_{i \mid s_v}, \qquad (i = 2, \ldots, k), \tag{4.45}$$

waarin $\pi_{i|s}$ de kans is op een juist antwoord gegeven dat de toetsscore gelijk is aan s. Het rechterlid van (4.44) is dus gelijk aan het rechterlid van (4.45), en deze gelijkheid geldt, ongeacht welke scores in de steekproef zijn geobserveerd. Daarom moet de gelijkheid ook term per term gelden, en we krijgen het belangrijke resultaat

$$\pi_{i/s} = \frac{\varepsilon_i \gamma^{(i)}_{s-1}(\boldsymbol{\varepsilon})}{\gamma_s(\boldsymbol{\varepsilon})}. \tag{4.46}$$

De oplossing van het stelsel (4.44) moet successief benaderd worden. Het zoeken van de oplossing is rekenintensief omdat veelvuldig de γ-functies moeten worden berekend. Een bijkomend probleem hierbij is dat bij het berekenen van die γ-functies, althans indien men er bepaalde algoritmen voor gebruikt, de resultaten zeer onnauwkeurig kunnen worden als gevolg van afrondingen. Om deze onnauwkeurigheden te vermijden, dient men algoritmen te gebruiken die nog meer tijd vergen. Deze omstandigheid brengt sommige auteurs er toe CML als schattingsmethode af te raden of zelfs af te wijzen (bijvoorbeeld Wainer & Mislevy, 1990, p. 80). Er is echter aangetoond dat met

een bepaalde berekeningsmethode van de symmetrische basisfuncties zeer nauwkeurige resultaten verkregen worden: bij k=5000 zijn slechts de laatste vier cijfers van het resultaat aangetast door afrondingsfouten (Verhelst, Glas & Van der Sluis, 1984). In gewone praktijktoepassingen waarbij k zelden groter is dan 100 is het verlies in de regel niet groter dan twee decimalen. In het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993) waar deze nauwkeurige methode is geïmplementeerd wordt gerekend met een nauwkeurigheid van ongeveer 14 decimalen, zodat van de berekende γ-functies de eerste 12 cijfers zeker correct zijn. Bovendien zijn de moderne computers zo snel dat het oplossen van (4.44) voor k=100 maar enkele minuten duurt. Praktische bezwaren tegen het gebruik van de CML-methode kunnen dus als volkomen achterhaald worden beschouwd. Voor technische details over het berekenen van de γ-functies en het oplossen van (4.44), zie Fischer (1974, hoofdstuk 14), Verhelst, Glas en van der Sluis (1984), Verhelst en Veldhuijzen (1991) en Verhelst, Glas en Verstralen (1993).

Met betrekking tot de statistische nauwkeurigheid van de schatters, moet het begrip informatie dat in paragraaf 4.2.1. werd besproken, uitgebreid worden tot het geval van meer parameters, waar men spreekt van een informatiematrix. Bij een model met k parameters is de informatiematrix een $k \times k$ symmetrische matrix $I(\beta)$, waarvan de cel *(i,j)* gegeven is door minus de verwachte waarde van de tweede partiële afgeleide van de log-aannemelijk- heidsfunctie naar de i-de en de j-de parameter. Voor de conditionele aannemelijkheidsfunctie (4.41) is dit dus

$$I_{ij}(\beta) = -\mathscr{E}\left[\frac{\partial^2 \ln L(\beta;\boldsymbol{X}\,|\,\boldsymbol{s})}{\partial\beta_i\,\partial\beta_j}\right]. \tag{4.47}$$

Toegepast op het Raschmodel geeft dit

$$I_{ij}(\beta) = \begin{cases} \sum_v [\pi_{i|s_v}(1-\pi_{i|s_v})] & \text{indien } i=j, \\ \sum_v [\pi_{ij|s_v} - \pi_{i|s_v}\pi_{j|s_v}] & \text{indien } i\neq j, \end{cases} \tag{4.48}$$

waarin

$$\pi_{ij|s_v} = P(X_{vi}=1, X_{vj}=1\,|\,s_v) = \frac{\varepsilon_i\varepsilon_j\gamma^{(i,j)}_{s_v-2}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}. \tag{4.49}$$

In (4.49) betekent *(i,j)* in superscript dat zowel ε_i als ε_j uit de argumentvector $\boldsymbol{\varepsilon}$ zijn weggelaten. De afleiding van (4.48) gebeurt geheel analoog aan de afleiding van (4.44). Details hierover zijn te vinden in Fischer (1974, p. 235 e.v.). De multivariate versie van het resultaat dat in 4.2.1. vermeld werd, luidt dan:

> De schatters van de *k*-1 vrije parameters zijn asymptotisch normaal verdeeld met als gemiddelde de werkelijke waarden van de parameters en de inverse van de informatiematrix als variantie-covariantie-matrix.

Net als in het univariate geval worden de itemparameters in (4.48) vervangen door hun ML-schattingen. De standaardfout (*SE*) van de itemparameterschatters is dan gegeven door de vierkantswortel van de elementen op de hoofddiagonaal van de inverse van $I(\boldsymbol{\beta})$.

In verband met de standaardfouten dient men zich te hoeden voor een veel voorkomende fout. Meestal wordt bij het rapporteren van de schattingen van de itemparameters, een standaardfout vermeld bij elk item. Dit betekent dat men een standaardfout krijgt voor *k* parameters, terwijl het model slechts *k*-1 vrije itemparameters heeft. Het antwoord op deze schijnbare paradox is dat de standaardfouten afhankelijk zijn van de gekozen normalisatie. Indien men bijvoorbeeld kiest $\beta_1 = 0$, dan is β_1 een constante en heeft per definitie een standaardfout van 0. De andere schattingen zullen een standaardfout opleveren die verschilt van 0. Gaan we nu over op een andere normalisatie, bijvoorbeeld met $\beta_2 = 0$, dan vinden we de nieuwe schattingen door van de eerste de oorspronkelijke schatting van β_2 af te trekken. Duiden we de nieuwe schattingen aan met $\hat{\tau}$, dan zijn de nieuwe schattingen en hun varianties gegeven in tabel 4.1

Tabel 4.1

Effecten van de normalisatie op schattingen en hun variantie

item	schatting bij $\beta_1 = 0$	schatting bij $\beta_2 = 0$	variantie bij $\beta_2 = 0$
1	0	$\hat{\tau}_1 = -\hat{\beta}_2$	$\text{var}(\hat{\tau}_1) = \text{var}(\hat{\beta}_2)$
2	$\hat{\beta}_2$	0	0
i (>2)	$\hat{\beta}_i$	$\hat{\tau}_i = \hat{\beta}_i - \hat{\beta}_2$	$\text{var}(\hat{\tau}_i) = \text{var}(\hat{\beta}_i) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_i,\hat{\beta}_2)$

Bij de veel gebruikte normalisatie waarbij de som van de schattingen gelijk is aan nul, beschouwt men *k* functies van de oorspronkelijke *k*-1 vrije parameters. Stel dat weerom de oorspronkelijke normalisatie gekozen was met $\beta_1 = 0$, dan zijn de *k* functies $\hat{\delta}_i$ waarvoor geldt dat $\sum_{i=1}^{k} \hat{\delta}_i = 0$ gegeven door

$$\hat{\delta}_i = \hat{\beta}_i - \frac{1}{k}\sum_{j=1}^{k}\hat{\beta}_j \tag{4.50}$$

en hun variantie is

$$\begin{aligned}\mathrm{var}(\hat{\delta}_i) &= \frac{(k-1)^2}{k^2}\mathrm{var}(\hat{\beta}_i) + \frac{1}{k^2}\sum_{j\neq i}\mathrm{var}(\hat{\beta}_j) \\ &\quad - \frac{2(k-1)}{k^2}\sum_{j\neq i}\mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j) + \frac{1}{k^2}\sum_{j\neq i}\sum_{\substack{m\neq i\\ m\neq j}}\mathrm{cov}(\hat{\beta}_j,\hat{\beta}_m),\end{aligned} \tag{4.51}$$

waarbij $\mathrm{var}(\hat{\beta}_1) = \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_i) = 0,\ (i\neq 1)$.

Het is instructief de CML-methode nog eens op een andere manier te bekijken. Voor een antwoordpatroon $\boldsymbol{x}$ met score s geldt

$$L(\beta,\theta;\boldsymbol{x},s) = P(\boldsymbol{x}|s)P(s|\theta). \tag{4.52}$$

De eerste factor in het rechterlid van (4.52) is de conditionele aannemelijkheidsfunctie gegeven door (4.40) en is onafhankelijk van θ. De tweede factor is de som van de kansen voor alle antwoordpatronen die score s opleveren, en is dus gegeven door

$$P(s|\theta) = \frac{\gamma_s(\boldsymbol{\varepsilon})\exp(s\theta)}{\prod_{i=1}^{k}[1+\varepsilon_i\exp(\theta)]}. \tag{4.53}$$

Deze kans is overduidelijk afhankelijk van θ maar ook van de itemparameters. Bij toepassing van CML wordt alleen de eerste factor in (4.52) gebruikt; de tweede factor wordt 'weggegooid'. Het lijkt er dus op dat door die tweede factor niet mee te nemen, informatie over de itemparameters, die in de score bevat is, wordt verwaarloosd, waardoor minder nauwkeurige schattingen van de itemparameters verkregen worden. Andersen (1970) heeft echter aangetoond dat dit niet zo is. De CML-methode gebruikt dus alle informatie over de itemparameters die in de gegevens aanwezig is.

Tot hiertoe is nog niets gezegd over de manier waarop de getoetste personen uit de populatie getrokken dienen te worden. Dit is met opzet gebeurd. Er is niet stilzwijgend verondersteld dat de steekproef een aselecte trekking moet zijn uit de populatie. Integendeel, door gebruik te maken van de CML-methode maakt het in principe niets uit hoe de steekproef uit de populatie is getrokken. Immers de CML-methode wordt gebruikt om iets te kunnen zeggen over de itemparameters en niet over de populatie

van personen. Bij de derde schat- tingsmethode, die in de volgende subparagraaf wordt besproken, hebben we dit voordeel niet. Dit voordeel van de CML-methode wordt vaak steekproefonafhankelijkheid genoemd. Als hierboven gezegd werd dat het 'in principe' niets uitmaakt hoe de steekproef wordt getrokken, wordt daarmee bedoeld dat CML niet in alle omstandigheden goed werkt. Als we bijvoorbeeld de gegevens inspecteren voor de analyse, en we gooien alle personen die item twee fout hadden uit de steekproef, dan zal de CML-methode geen consistente schatters van de itemparameters opleveren. Wanneer het precies wel en niet goed gaat, wordt gedetailleerd uiteengezet in hoofdstuk 6. Een tweede kanttekening die bij de notie van steekproefonafhankelijkheid gemaakt moet worden betreft de nauwkeurigheid van de parameterschattingen. Twee steekproeven van dezelfde omvang leveren niet noodzakelijkerwijze even nauwkeurige schattingen van de parameters op. In paragraaf 4.2.5 wordt hierop teruggekomen.

4.2.4 MML-schatting in het Raschmodel

Een tweede methode om de individuele θ-parameters kwijt te raken bestaat eruit ze een andere status te geven. De status van de θ-waarden is het standpunt van waaruit men de gegevens beschouwt. Tot nog toe hebben we eigenlijk impliciet aangenomen dat, als Jan en Piet tot de steekproef behoren, we ter zelfder tijd geïnteresseerd zijn in de waarde van de itemparameters en in de θ-waarde van Jan en Piet en van alle andere personen die tot de steekproef behoren. Een ander standpunt is dat het ons eigenlijk niet kan schelen wie er in de steekproef zit, omdat we alleen maar geïnteresseerd zijn in de itemparameters. Dit impliceert dat we de steekproef als een aselecte steekproef uit een of andere populatie beschouwen, en dat we de gedragingen van die toevallige steekproef willen gebruiken om de itemparameters te schatten. Dit standpunt biedt de mogelijkheid om θ kwijt te raken op de volgende manier.

Veronderstel dat θ slechts drie verschillende waarden kan aannemen in de populatie, namelijk -1, 0 en 1, en veronderstel dat deze waarden in de populatie voorkomen met een proportie van respectievelijk .25, .35 en .40. We beschouwen nu de kans dat we het ant-woordpatroon $\boldsymbol{x}$ = (1 0 1) observeren bij aselecte trekking van een persoon uit de populatie. Deze kans is gegeven door

$$P(\boldsymbol{x}) = 0.25 \times P(\boldsymbol{x}|\theta = -1) + 0.35 \times P(\boldsymbol{x}|\theta = 0) + 0.40 \times P(\boldsymbol{x}|\theta = 1).$$

Dat wil zeggen, als we θ niet kennen, kunnen we alle conditionele kansen $P(\boldsymbol{x}|\theta)$ als het ware gaan middelen door te vermenigvuldigen met de kans dat die θ optreedt, en die gewogen conditionele kansen op te tellen. Het resultaat noemt men marginale kans. Vandaar de eerste M in MML. Laten we dit nu veralgemenen tot de situatie waarin het aantal verschillende waarden dat θ kan aannemen gelijk is aan W:

$$P(\boldsymbol{x}) = \sum_{j=1}^{W} P(\boldsymbol{x}|\theta_j)\, P(\theta_j). \tag{4.54}$$

Het gebruik van (4.54) zonder meer is niet erg aantrekkelijk, omdat we dan een waarde voor W moeten kennen, de verschillende waarden die θ kan aannemen en de kansen $P(\theta_j)$. Als we die niet kennen, moeten we ze ook uit de data schatten, zodat er naast de itemparameters nog eens $2W$ parameters bijkomen: W waarden van θ, W-1 vrije kansen $P(\theta_j)$ en W zelf. Boven- dien is W discreet, en kan bijgevolg niet geschat worden met de standaardmethodes die in paragraaf 4.2.1 zijn uiteengezet. Het gebruik van het rechterlid van (4.54) als aannemelijk- heidsfunctie brengt dan ook enkele moeilijke problemen met zich mee. Voor enkele interessante resultaten bij deze benadering, zie De Leeuw en Verhelst (1986), Follman (1988) en Lindsay, Clifford en Grego (1991).

Hoe paradoxaal het ook klinkt, het probleem wordt veel eenvoudiger door θ oneindig veel waarden te laten aannemen, en nog sterker: door θ continu te laten worden, en een bepaalde regel te veronderstellen waaruit de 'kans' op een bepaalde θ uit θ zelf bepaald kan worden. We mogen bij continue variabelen niet meer spreken van kans; men spreekt van dichtheid. Die dichtheid duiden we aan met het functiesymbool g. We kennen een heel populaire dichtheid, namelijk die van de normale verdeling:

$$g(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right], \tag{4.55}$$

waarin $\pi = 3.14159...$ We zien dat in die functieregel twee parameters voorkomen, namelijk μ en σ^2, het gemiddelde en de variantie van θ. De marginale kans van antwoordpatroon $\boldsymbol{x}$ in het geval we een normale verdeling veronderstellen van θ, is gegeven door

$$\begin{aligned} P(\boldsymbol{x}) &= \int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)\, g(\theta)\, d\theta \\ &= \int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)\, \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta. \end{aligned} \tag{4.56}$$

Formule (4.56) is niet meer afhankelijk van θ, want die is er uitgeïntegreerd, wel van de itemparameters en van de twee verdelingsparameters μ en σ^2. Indien we deze marginale kans nu beschouwen als functie van die parameters, dan krijgen we de marginale aannemelijkheidsfunctie voor het antwoordpatroon $\boldsymbol{x}$. De aannemelijkheidsfunctie voor alle geobserveerde antwoordpatronen samen is dan gegeven door

$$L(\beta,\mu,\sigma^2;\boldsymbol{X}) = \prod_{v=1}^{n} \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta . \tag{4.57}$$

Nemen we hiervan de logaritme,

$$\ln L(\beta,\mu,\sigma^2;\boldsymbol{X}) = \sum_{v=1}^{n} \ln \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta , \tag{4.58}$$

dan stuiten we op de moeilijkheid dat we de logaritme van een integraal moeten nemen. Zulke uitdrukkingen laten zich in de regel niet vereenvoudigen, tenzij men een expliciete uitdrukking kan vinden voor de integraal, dat wil zeggen een uitdrukking zonder integraal- teken. Niemand echter heeft zo'n expliciete uitdrukking gevonden, en waarschijnlijk bestaat die zelfs niet. De uitdrukking in het rechterlid van (4.58) kan dan ook niet teruggebracht worden tot de standaarduitdrukking voor de exponentiële familie, en er kan dus geen beroep gedaan worden op de eigenschappen van de exponentiële familie. Het vinden van het maximum van (4.58) is dan ook geen eenvoudige aangelegenheid. Op de verdere details van dit probleem gaan we niet in. Er zijn verschillende computerprogramma's in de handel die MML-schattingen berekenen, en ook de bijbehorende standaardfouten. Bijvoorbeeld BILOG (Mislevy & Bock, 1986), MULTILOG (Thissen, 1988) en het reeds eerder vermelde OPLM. In de statistiek is bewezen (Kiefer & Wolfowitz, 1956) dat door deze methode consistente schattingen van alle parameters worden verkregen.

We sluiten deze paragraaf af met een korte vergelijking van de CML- en de MML-methode. Het belangrijkste verschil tussen beide methodes bestaat erin dat bij CML geen enkele veronderstelling wordt gemaakt over de verdeling van θ in de populatie, terwijl dat bij MML wel wordt gedaan. Het is bij MML helemaal niet noodzakelijk een normale verdeling te veronderstellen. Men zou ook een andere verdeling kunnen aannemen, zie bijvoorbeeld Andersen en Madsen (1977). Belangrijk is echter in te zien dat de veronderstelling over de verdeling nu deel gaat uitmaken van het model. Dus

als we MML toepassen, dan vermengen we als het ware twee modellen: het Raschmodel dat iets vertelt over de antwoorden gegeven θ, en de normale verdeling die vertelt hoe de θ's in de populatie zijn verdeeld. De verstrengeling van beide modellen gebeurt op een heel diep niveau (zie formule (4.56)), zodanig dat beide onderdelen niet eenvoudig uit elkaar zijn te halen. Maken we een fout in de veronderstelling over de normale verdeling, hetzij omdat θ niet normaal verdeeld is, hetzij omdat de steekproef niet aselect uit de normale verdeling is getrokken, dan heeft dat als gevolg dat er ook systematische fouten geïntroduceerd worden in de schatting van de itemparameters. Een gebruiker die MML gebruikt stelt zich dus iets kwetsbaarder op.

Het voordeel van MML is wel dat de verdelingsparameters gelijktijdig met de itemparameters geschat kunnen worden. Indien men in beide geïnteresseerd is, is MML de meest efficiënte methode. In paragraaf 4.4 en uitvoeriger in hoofdstuk 6, waar onvolledige designs worden besproken, zullen we zien dat in sommige omstandigheden CML helemaal niet kan toegepast worden, maar MML wel.

4.2.5 Een voorbeeld

Een goede manier om een indruk te krijgen van de eigenschappen van schattingen is het analyseren van artificiële of gesimuleerde data. Immers, indien we reële data analyseren, weten we nooit of aan de veronderstellingen van het model is voldaan, en bovendien kennen we de echte waarden van de parameters niet. Artificiële data zijn afkomstig van een computerprogramma dat geïnstrueerd kan worden zich volgens het model te gedragen. Essentieel daarbij is dat er een programma voorhanden is dat een aselecte trekking uit een bepaalde verdeling kan uitvoeren. Zulke programma's bestaan en zijn uitvoerig in de statistische literatuur beschreven.

Stel dat we een antwoordpatroon willen genereren van een artificieel persoon die aselect uit de standaardnormale verdeling is getrokken. De toets bestaat uit k=3 items die aan het Raschmodel voldoen en parameterwaarden hebben van respectievelijk -1, 0 en 1. Het pro- gramma start met het trekken van een θ-waarde uit de standaardnormale verdeling. Neem aan dat θ = 0.2. Dan kan berekend worden met behulp van (4.5) dat

$$f_1(0.2) = 0.769, \quad f_2(0.2) = 0.550, \quad f_3(0.2) = 0.310.$$

Vervolgens wordt uit de uniforme verdeling op het interval (0,1) een toevalsgetal p_1 getrokken. Voor de toevalsvariabele p_1 geldt dus dat

$$P(p_1 \leq x) = x, \quad (0 < x \leq 1)$$

en dus $P(p_1 \leq 0.769) = 0.769$. Indien $p_1 \leq 0.769$, krijgt de toevalsvariabele X_1, het antwoord op item 1, de waarde 1, anders 0. Deze procedure wordt herhaald voor elk item, waarbij voor elk item i dus een nieuw en onafhankelijk toevalsgetal p_i uit de uniforme verdeling wordt getrokken. Voor de getrokken waarde van θ is de antwoordregel dus gegeven door

$$X_i = \begin{cases} 1 \text{ indien } p_i \leq f_i(\theta), \\ 0 \text{ indien } p_i > f_i(\theta). \end{cases}$$

De hele hierboven beschreven procedure wordt herhaald voor elk van de n artificiële personen.

In tabel 4.2 staan de resultaten van een analyse op artificiële data, met $n = 500$ personen aselect getrokken uit de standaardnormale verdeling. Het aantal items is acht en de itemparameters zijn -2, -1.5, -1, -0.5, 0.5, 1, 1.5 en 2.

Tabel 4.2
Parameterschattingen uit artificiële data

	CML met $\Sigma_i \hat{\delta}_i = 0$		CML met $\hat{\beta}_1 = 0$		CML met $\hat{\tau}_2 = 0$		MML met $\Sigma_i \hat{\delta}_i = 0$	
β_i	$\hat{\delta}_i$	SE($\hat{\delta}_i$)	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	$\hat{\tau}_2$	SE($\hat{\tau}_2$)	$\hat{\delta}_i$	SE($\hat{\delta}_i$)
-2.	-2.239	0.133	0	---	-0.724	0.181	-2.264	0.135
-1.5	-1.515	0.111	0.724	0.181	0	---	-1.511	0.113
-1.	-1.073	0.103	1.166	0.177	0.441	0.158	-1.063	0.104
-0.5	-0.283	0.096	1.956	0.175	1.231	0.154	-0.273	0.095
0.5	0.609	0.098	2.848	0.180	2.123	0.159	0.615	0.097
1.	0.924	0.101	3.163	0.183	2.439	0.162	0.930	0.101
1.5	1.560	0.113	3.799	0.193	3.075	0.174	1.561	0.113
2.	2.018	0.128	4.257	0.205	3.533	0.187	2.004	0.125

Voor de werkelijke parameters wordt het symbool β gebruikt, voor de CML-schattingen en de MML-schattingen waarvoor de som van de schattingen gelijk is aan 0, wordt het symbool $\hat{\delta}_i$ gebruikt. Voor de CML-schattingen waarbij de parameterschatting van het eerste item gelijk gesteld is aan 0 gebruiken we het symbool $\hat{\beta}_i$ en voor de schattingen waar de parameter van het tweede item gelijkgesteld is aan 0 wordt $\hat{\tau}_i$ gebruikt. Dit

is in overeenstemming met de notatie die gebruikt is in paragraaf 4.2.3 bij de discussie over de standaardfouten. Uit tabel 4.2 zijn enkele interessante bevindingen af te leiden.

Voor de CML-schattingen en de MML-schattingen met dezelfde normering krijgen we ongeveer dezelfde uitkomsten. In alle gevallen ligt de ware parameter binnen het 95%- betrouwbaarheidsinterval rond de geschatte waarde. Ook de geschatte standaardfouten zijn ongeveer aan elkaar gelijk. Indien men de nauwkeurigheid van de schattingen onvoldoende vindt, dan kan de nauwkeurigheid opgevoerd worden door de steekproef groter te maken. Uit (4.48) volgt dat elke persoon een eigen onafhankelijke bijdrage heeft aan de informatiematrix. Nemen we de steekproef dubbel zo groot, dan verdubbelt ook de informatie, en de variantie van de schatters wordt gehalveerd. De standaardfout neemt dus af met een factor $\sqrt{2}$. Wil men de standaardfouten halveren, dan dient men dus een steekproef te nemen die vier maal zo groot is. Dit geldt zowel voor MML als voor CML.

De drie gerapporteerde CML-schattingen verschillen slechts een constante van elkaar, zoals kan afgeleid worden uit tabel 4.2. Normeren door één parameter gelijk aan 0 of een andere constante te stellen, resulteert in veel grotere standaardfouten voor de andere parameters dan in het geval dat de som van de schattingen gelijk wordt gesteld aan 0. Als voorbeeld berekenen we de correlatie tussen $\hat{\beta}_2$ en $\hat{\beta}_3$. Passen we de formule rechtsonder in tabel 4.1 toe voor *i*=3, dan vinden we

$$0.158^2 = 0.177^2 + 0.181^2 - 2\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3)$$

waaruit volgt dat $\mathrm{cov}(1\hat{\beta}_2, \hat{\beta}_{31}) = 0.01956$. De correlatie tussen de schatters bedraagt dus

$$\mathrm{corr}(\hat{\beta}_2, \hat{\beta}_3) = \frac{0.01956}{0.181 \times 0.177} = 0.611.$$

De hoogte van de correlatie is niet afhankelijk van de steekproefgrootte, noch van het aantal items, maar wel van de informatie die verkregen wordt over het item waarop genormeerd wordt, dit is het desbetreffende element op de hoofddiagonaal van de informatiematrix; zie formule (4.48). Is deze informatie relatief laag, dan zal de correlatie hoger uitvallen dan wanneer die informatie relatief groot is. Dit wordt geïllustreerd in tabel 4.3. De resultaten in de tweede en derde kolom van deze tabel hebben betrekking op dezelfde gegevens als tabel 4.2; in de vierde en vijfde kolom staan de resultaten op een onafhankelijke steekproef ter grootte van 5000, maar met dezelfde itemparameters als in het eerste voorbeeld.
De twee steekproefgroottes leveren nagenoeg dezelfde schatting van de correlaties op. Merk op dat de getallen in de vierde kolom van tabel 4.3 ongeveer tien maal zo groot

zijn als de getallen in de tweede kolom: de informatie neemt evenredig toe met het aantal observaties. De verhouding is niet exact 10, omdat de kolommen alleen schattingen van de informatie bevatten.

Tabel 4.3

Correlatie tussen CML-schatters als functie van het item waarop genormeerd wordt

	n=500		n=5000	
item	informatie	$\text{corr}(\hat{\beta}_2,\hat{\beta}_3)$	informatie	$\text{corr}(\hat{\beta}_2,\hat{\beta}_3)$
1	47.3	.611	490	.588
4	84.4	.452	836	.450
5	86.6	.461	828	.472
6	77.8	.499	749	.507
7	61.0	.569	633	.557
8	52.7	.607	505	.615

De standaardfouten in tabel 4.2 zijn niet voor alle items even groot. Dit hangt eveneens samen met de informatie die de gegevens over het item opleveren. De iteminformatiefunctie wordt gegeven door de elementen op de diagonaal van de informatiematrix, zie (4.48). In veel toepassingen wordt alleen van deze elementen gebruik gemaakt om een schatting van de standaardfout te maken:

$$SE^{*}(\hat{\beta}_i) = [I_{ii}(\hat{\beta})]^{-\frac{1}{2}} = \left[\sum_v \hat{\pi}_{i|s_v}(1-\hat{\pi}_{i|s_v})\right]^{-\frac{1}{2}}. \qquad (4.59)$$

De uitkomst van (4.59) kan dus om drie redenen van de echte standaardfout verschillen. Ten eerste, in het rechterlid worden niet de echte conditionele kansen $\pi_{i|s}$, maar schattingen ingevuld. Ten tweede wordt een asymptotisch resultaat toegepast op een eindige steekproef en ten derde worden de buitendiagonale elementen van de informatiematrix verwaarloosd. Toch wordt (4.59) in de praktijk vaak gebruikt, en soms niet terecht zoals we verderop zullen zien. De eenvoudige structuur van (4.59) laat ons echter twee zaken duidelijk zien. In de eerste plaats het effect van de steekproefgrootte op de nauwkeurigheid van de schattingen. Elke antwoordpatroon in de steekproef levert precies één term aan de som in het rechterlid van (4.59). Als we uit een bepaalde populatie twee aselecte steekproeven trekken, de eerste van *n* personen, en de tweede van 2*n* personen, dan zal de som in het rechterlid van (4.59) voor de tweede steekproef ongeveer twee keer zo groot zijn als voor de eerste steekproef, en dus zullen de standaardfouten van de itemparemeterschattingen in de eerste steekproef ongeveer $\sqrt{2}$

zo groot zijn als in de tweede steekproef. In de tweede plaats kunnen we (4.59) gebruiken om te laten zien dat we voorzichtig moeten omspringen met het theoretische voordeel van de steekproefonafhankelijkheid. De maximale waarde van het produkt $\pi_{i|s}(1-\pi_{i|s})$ bedraagt 0.25 en wordt bereikt indien $\pi_{i|s}$=0.5. Indien de score 0 is of *k*, is de bijdrage aan de informatie precies 0. Stel nu dat we aan een steekproef van n personen een toets voorleggen die veel te moeilijk is, zodat relatief veel personen een score 0 behalen. De antwoordpatronen van deze personen dragen dus niets bij aan de iteminformatie, en de standaardfouten van de parameterschattingen zullen groter zijn dan in het geval van een even grote steekproef waarbij de moeilijkheidsgraad van de items goed overeenkomt met de vaardigheid van de personen. Het voordeel van steekproefonafhankelijkheid moet dus niet gebruikt worden om een toets voor te leggen aan een willekeurige verzameling personen. In hoofdstuk 7 zullen we twee voorbeelden zien waarbij op een verstandige manier voordeel is gehaald uit de steekproefonafhankelijkheid van de CML-schatters.

De reden waarom (4.59) in de praktijk vaak gebruikt wordt, is dat het uitrekenen en inverteren van de hele informatiematrix erg tijdrovend wordt indien het aantal items groot is. Formule (4.59) kan ook gebruikt worden voor het item waarop genormeerd is. Gaat men naderhand de oplossing centreren, dan wordt ook dezelfde formule gebruikt om de standaardfout van de *k* parameters te berekenen. Standaardfouten kunnen dus op veel verschillende manieren geschat worden, en de resultaten kunnen nogal uiteenlopen. Dit kunnen we zien door een speciaal geval te bestuderen. Veronderstel dat de parameters van de *k* items in een toets allemaal aan elkaar gelijk zijn. De informatie die we over elk item inwinnen zal dus ook dezelfde zijn voor alle items. De elementen op de hoofddiagonaal van de informatiematrix zullen dus ook aan elkaar gelijk zijn. Veronderstel dat die informatie gelijk is aan c^2. De waarde van c^2 is afhankelijk van de grootte van de steekproef en van de moeilijkheid van de items in vergelijking met de gemiddelde vaardigheid. In tabel 4.4 worden de asymptotische standaardfouten, berekend uit de inverse van de informatiematrix, en hun schattingen, gebaseerd op formule (4.59), gegeven.

Tabel 4.4
Standaardfout bij *k* items van dezelfde moeilijkheid

normering	*SE*	formule (4.59)
op één item	$\frac{1}{c}\sqrt{\frac{2(k-1)}{k}}$	$\frac{1}{c}$

gecentreerd	$\frac{1}{c}\frac{k-1}{k}$	$\frac{1}{c}$

Voor de theoretische afleiding van dit resultaat, verwijzen we naar Verhelst (1993). Voor de gecentreerde oplossing is (4.59) dus een goede benadering indien het aantal items niet te klein is. Merk ook op dat (4.59) systematisch de standaardfout overschat. Kiezen we echter een oplossing waarbij op één item genormeerd is, dan geeft (4.59) een grove onderschatting van de standaardfout: het effect van de verwaarlozing van de buitendiagonale elementen van de informatiematrix komt dus ongeveer overeen met het overwaarderen van de steekproefgrootte met een factor 2. De gecentreerde oplossing verdient dus de voorkeur. Tenslotte is het gemakkelijk te controleren dat de correcte standaardfout bij een gecentreerde oplossing kleiner is dan bij normering op één item.

4.3 Het toetsen van het Raschmodel

In paragraaf 4.2.1 hebben we de grootste-aannemelijkheidsschatter voor de parameter van een muntstuk opgesteld. Daar bleek dat we enkel het relatief aantal successen hoefden te kennen om die parameter te schatten. De observaties kunnen ons niet méér informatie opleveren. Indien we zeker zouden zijn dat aan de veronderstellingen van het model was voldaan, dan hoefden we ook niets meer te weten. Bij het opgooien van een muntstuk zijn die veronderstellingen eenvoudig: de kans op succes moet onveranderd blijven en de uitkomst bij elke worp moet onafhankelijk zijn van de uitkomsten van de andere worpen. Veronderstel nu dat het opgooien zo klungelig gebeurt dat de uitkomst bij een bepaalde worp bijna zeker gelijk is aan de uitkomst van de vorige worp, bijvoorbeeld omdat het muntstuk maar een heel klein beetje wordt opgetild en dan weer losgelaten. Als die afhankelijkheid heel sterk is, is het mogelijk dat bij 100 keer opgooien het muntstuk 99 keer op munt valt, ook al is het niet vervalst. We kunnen dan nog wel de techniek van het schatten gaan toepassen, doch de conclusie dat het muntstuk onzuiver is, is niet terecht, omdat niet voldaan is aan de veronderstellingen van het model. Om na te gaan of aan de veronderstellingen van het model is voldaan, kunnen we natuurlijk de experimentele procedure aan een nader onderzoek onderwerpen. Indien het muntstukexperiment is uitgevoerd zoals hierboven beschreven, zullen we niet geneigd zijn de resultaten serieus te nemen. Indien bij de dataverzameling van toetsgegevens de afname niet serieus gebeurt, bijvoorbeeld omdat de leerlingen alle gelegenheid krijgen elkaar te consulteren bij het beantwoorden van de items,

kunnen we beter de statistische verwerking achterwege laten, want de belangrijke eis van experimentele onafhankelijkheid is geschonden, en alle conclusies die uit een statistische analyse volgen, berusten op drijfzand. Echter, een zorgvuldige dataverzameling is wel een noodzakelijke, doch geen voldoende voorwaarde opdat alle veronderstellingen van het model vervuld zijn. De reden hiervoor is dat schendingen van het model erg subtiel kunnen zijn. De algemene strategie om modelschendingen te ontdekken is het nauwkeurig onderzoeken van de fijnere structuur van de data, met name die aspecten van de data die niet zijn gebruikt om de parameters te schatten.

We beginnen met een voorbeeld uit het muntexperiment. Indien we 100 keer een zuivere munt opgooien, en we stellen vast dat het muntstuk de eerste 50 keer munt valt, en vervolgens 50 keer kruis, dan zullen we het muntstuk of het experiment niet vertrouwen. We verwachten dat de afwisseling 'k-m' of 'm-k' meer dan één keer optreedt. Observeren we echter een volmaakte regelmaat waarbij k en m elkaar iedere keer weer afwisselen, dan is dit ook verdacht. Het aantal afwisselingen mag dus niet te groot zijn maar ook niet te klein. De statistische theorie wordt gebruikt om precies aan te geven wat bedoeld wordt met te groot of te klein. De toetsingsprocedure voor dit probleem staat beschreven in Siegel en Castellan (1988, p. 58-64).

In het Raschmodel is de ruwe score, het aantal items juist, een voldoende steekproefgrootheid voor de latente variabele θ. In paragraaf 4.5 zullen we zien dat iedereen met dezelfde score ook dezelfde schatting van θ krijgt, ongeacht welke items juist beantwoord zijn. Dit betekent echter niet dat bij de subgroep van personen die dezelfde score hebben alle mogelijke antwoordpatronen even waarschijnlijk zijn. Beschouwen we een simpel voorbeeld. Laat de toets bestaan uit $2k$ items, waarvan k gemakkelijke met itemparameter $\varepsilon_i = 2$, $i = 1,\ldots,k$ en k moeilijke met itemparameter $\varepsilon_i = 0.5$, $i = k+1,\ldots,2k$ (zie formule (4.37)). In de subpopulatie van personen die precies k items juist hebben, is de kans dat de k gemakkelijkste items juist zijn beantwoord, gegeven door

$$P(X_1 = \ldots = X_k = 1\,,\, X_{k+1} = \ldots = X_{2k} = 0 \,|\, s = k) \;=\; \frac{2^k}{\gamma_k(\boldsymbol{\varepsilon})},$$

en de kans dat de k moeilijkste juist zijn is

$$P(X_1 = \ldots = X_k = 0\,,\, X_{k+1} = \ldots = X_{2k} = 1 \,|\, s = k) \;=\; \frac{2^{-k}}{\gamma_k(\boldsymbol{\varepsilon})}\,.$$

De verhouding tussen die twee kansen is 2^{2k}. Bij 10 items en een score van 5 verwachten we dus 1024 keer zoveel respondenten die de vijf makkelijkste items juist hebben als respon- denten met de vijf moeilijkste items juist. Indien we in een

steekproef ongeveer gelijke aantallen zouden vinden, is dat een voldoende reden om de geldigheid van het model in twijfel te trekken. Dit voorbeeld maakt ook duidelijk dat een theorie die geen absolute uitspraken doet over het gedrag wel degelijk gefalsificeerd kan worden. De kans op een juist antwoord in het Raschmodel is altijd strikt groter dan 0 en strikt kleiner dan 1 ongeacht de waarde van θ. Hoewel dus met elke θ-waarde alle antwoordpatronen mogelijk zijn, zijn ze niet allemaal even waarschijnlijk, en deze ongelijke waarschijnlijkheden dienen weerspiegeld te worden in ongelijke relatieve frequenties in de steekproef. De statistische theorie wordt gebruikt om aan te geven hoe nauwkeurig die weerspiegeling dient te zijn.

4.3.1 De veronderstellingen van het Raschmodel

In paragraaf 4.1 is gezegd dat een belangrijke reden om het Raschmodel als meetmodel te kiezen wiskundige elegantie is. Dit is ongetwijfeld waar, maar men kan zich de vraag stellen of er geen andere modellen bestaan die wiskundig even elegant zijn, en toch drastisch van het Raschmodel verschillen. In de literatuur zijn verschillende pogingen ondernomen om het Raschmodel af te leiden uit een aantal eenvoudige aannames. Deze aannames worden ook axioma's genoemd. Het is mogelijk het Raschmodel af te leiden uit verschillende ver- zamelingen van axioma's. Voor een overzicht, zie Fischer (in voorbereiding). Wij zullen één stel aannames bespreken, zonder echter de afleiding aan te tonen, omdat deze wiskundig nogal moeilijk is. Deze aannames zijn:

(1) de itemresponscurve $f_i(\theta)$ is continue en strikt stijgend voor alle waarden van θ en voor alle items *i* in de beschouwde itemverzameling. θ is een unidimensionale grootheid en kan een willekeurige reële waarde aannemen;

(2) voor alle items *i* zijn de limieten (4.7) geldig:

$$\lim_{\theta\to\infty} f_i(\theta) = 1, \qquad \lim_{\theta\to-\infty} f_i(\theta) = 0;$$

(3) het axioma van de lokale stochastische onafhankelijkheid is geldig;

(4) de ruwe score $s = \sum_i x_i$ is een voldoende steekproefgrootheid voor θ.

Er kan mathematisch worden aangetoond dat de vier bovenstaande axioma's equivalent zijn met het Raschmodel. Voor de praktijk betekent dit dat schending van één of meer van die axioma's automatisch een schending is van het Raschmodel.

Het eenvoudigste voorbeeld van een schending wordt wellicht gegeven door het gebruik van meerkeuzevragen: uit axioma (2) volgt dat de kans op een juist antwoord, gegeven dat de vaardigheid zeer klein is ($\theta\to-\infty$), praktisch gelijk moet zijn aan 0. Indien er in zo'n geval geraden wordt tussen bijvoorbeeld vier alternatieven, is de kans op een juist antwoord 0.25. Voor dit soort items geeft het Raschmodel dus geen juiste

beschrijving. In de praktijk betekent dit dus dat raadgedrag een oorzaak kan zijn van de ongeldigheid van het Raschmodel. In paragraaf 7.2 wordt uitvoerig op dit probleem ingegaan.

Een tweede soort inbreuk die voor de praktijk relevant is, wordt gewoonlijk aangeduid met het begrip multidimensionaliteit. In axioma (1) is sprake van een unidimensionale grootheid θ. Neem aan dat θ staat voor numerieke vaardigheid. Veronderstel verder dat de items bestaan uit een aantal redaktiesommen, die een beroep doen zowel op deze numerieke vaardigheid als op verbale vaardigheid. Het is zeer wel mogelijk dat aan axioma (1) voldaan is, doch beschouwen we nu tegelijkertijd axioma (3). Dit axioma impliceert dat, indien θ constant wordt gehouden, de covariantie tussen alle itemantwoorden 0 is. Als numerieke vaardigheid en verbale vaardigheid niet precies hetzelfde betekenen, is het natuurlijk zo dat in een subpopulatie waar θ constant is, er nog variabiliteit in de verbale vaardigheid zal overblijven, en omdat we aangenomen hebben dat het antwoord gedeeltelijk door de verbale vaardigheid wordt bepaald, zal de covariantie tussen de itemantwoorden niet 0 zijn. Samenvattend kunnen we dus stellen dat, indien de items een beroep doen op meerdere vaardigheden die niet perfect correleren, en θ verwijst naar één van die vaardigheden, dan is automatisch het axioma van de lokale stochastische onafhankelijkheid geschonden. Door het hierboven gegeven voorbeeld iets aan te scherpen is ook duidelijk te zien dat het vierde axioma geschonden is. Veronderstel dat de helft van de items uitsluitend een beroep doen op verbale vaardigheid, en de andere helft uitsluitend op numerieke vaardigheid. Veronderstel bovendien dat verbale en numeriek vaardigheid in de populatie zeer laag correleren. Beschouw nu twee personen, A en B, die beide de helft van de items juist beantwoorden: persoon A heeft uitsluitend de verbale items juist en persoon B uitsluitend de numerieke items. Hoewel beide personen dezelfde ruwe score hebben behaald, ligt het voor de hand de numerieke vaardigheid, θ, van persoon B hoger in te schatten dan die van persoon A, doch dit is hetzelfde als het verwerpen van axioma (4).

Uit het voorgaande mag niet worden afgeleid dat rekenitems alleen aan het Raschmodel voldoen, indien ze uitsluitend een beroep doen op numerieke vaardigheid en niet op verbale vaardigheid. IRT-modellen zijn wiskundige modellen die voorspellingen doen over het gedrag van personen die de items beantwoorden. Indien deze voorspellingen juist zijn kan men daaraan het argument ontlenen dat de items in de toets een unidimensionale vaardigheid meten. Of deze vaardigheid een numerieke dan wel een mengsel van numerieke en verbale vaardigheden is, is een kwestie van interpretatie.

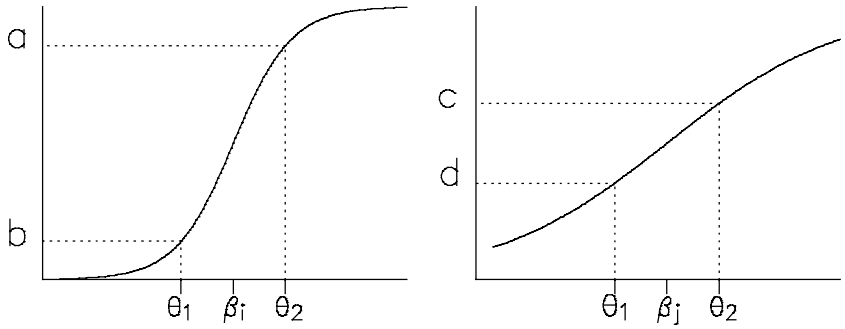
Voor een goed begrip van axioma (4) introduceren we een IRT-model dat algemener is dan het Raschmodel, namelijk het twee parameter logistisch model, dat ook wel

aangeduid wordt als het Birnbaummodel (Birnbaum, 1968). Het onderscheid tussen het Raschmodel en het Birnbaummodel hangt nauw samen met een eigenschap van het Raschmodel die uitgebeeld is in figuur 4.3: de curven van twee itemresponsfuncties snijden elkaar nooit.

Beschouw nu figuur 4.6. Daarin zijn twee itemresponscurves afgebeeld voor de items *i* en *j*. We nemen aan dat $\beta_i = \beta_j$. Beschouw nu twee personen met latente vaardigheden θ_1 en θ_2. Uit de figuur is duidelijk dat

$$f_i(\theta_2) - f_i(\theta_1) = a - b > c - d = f_j(\theta_2) - f_j(\theta_1).$$

Dit betekent dat we op grond van item *i* een beter onderscheid kunnen maken tussen die twee personen dan op grond van item *j*, hoewel beide items even moeilijk zijn. Anders gezegd: item *i* discrimineert beter dan item *j*. Dit betere discriminerend vermogen komt tot uiting in het steilere verloop van de itemresponscurve van item *i*. Merk overigens op dat dit discriminerend vermogen een plaatselijke eigenschap is: twee personen met een verschillende vaardigheid die voor beiden veel groter is dan de moeilijkheidsgraad van het item, zullen beiden bijna zeker het item oplossen en dus kan het item geen onderscheid maken tussen beider vaardigheid. Het discriminerend vermogen van een item wordt dus afgemeten aan de snelheid waarmee de itemresponsfunctie verandert in de buurt van de moeilijkheidsparameter.



Figuur 4.6
Twee items die verschillend discrimineren

Als we binnen de familie van logistische functies blijven, kunnen we dit verschil in discriminerend vermogen uitdrukken door een iets gecompliceerder functievorm te kiezen dan in het Raschmodel. De formule voor functies zoals weergegeven in figuur 4.6 is:

$$P(X_i=1|\theta) = f_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (a_i>0). \tag{4.60}$$

De grootheid a_i wordt de discriminatieparameter van het item genoemd. Een item wordt dus gekenmerkt door twee parameters: een moeilijkheidsparameter β_i en een (positieve) discriminatieparameter a_i. Merk op dat formule (4.60) onveranderd blijft indien we zowel θ als β_i met een willekeurige positieve constante c vermenigvuldigen en a_i door c delen. Vergelijken we nu (4.60) met (4.5), dan zien we dat in (4.5) het verschil θ-β_i met 1 is vermenigvuldigd. Formule (4.5) is dus een speciaal geval van (4.60), waarbij $a_i = 1$ voor alle items i. Maar omdat we de discriminatieparameters met een willekeurige positieve constante mogen vermenigvuldigen, kunnen we zeggen dat het Raschmodel een speciaal geval is van het Birnbaummodel waarbij alle discriminatieparameters aan elkaar gelijk zijn. In figuur 4.6 heeft item i een grotere discriminatieparameter dan item j.

Voor een antwoordpatroon $\boldsymbol{x}$ is met gebruik van (4.60) gemakkelijk aan te tonen dat de log-aannemelijkheidsfunctie gegeven is door

$$\ln L(\beta,\boldsymbol{a},\theta;\boldsymbol{x}) = \theta\sum_i a_i x_i - \sum_i x_i a_i \beta_i - \sum_i \ln\{1+\exp[a_i(\theta-\beta_i]\}, \qquad (4.61)$$

waaruit duidelijk blijkt dat de gewone somscore geen voldoende steekproefgrootheid is voor θ. In het Birnbaummodel is dus niet voldaan aan axioma (4). De gewogen somscore $\sum_i a_i x_i$ is wel voldoende, doch deze grootheid is een functie van de onbekende discriminatieparameters. Het Birnbaummodel behoort dus ook niet tot de exponentiële familie. Nadere beschouwingen over dit model worden in hoofdstuk 5 gegeven.

Naast de axioma's (1) tot (4) zijn er nog een paar veronderstellingen die strikt genomen geen axioma's zijn, doch die men zou kunnen omschrijven als algemene voorwaarden die vervuld moeten worden om het model te kunnen toepassen en toetsen. De eerste voorwaarde, die reeds ter sprake kwam, is experimentele onafhankelijkheid bij de dataverzameling. Indien niet aan die voorwaarde voldaan is, snijden we onszelf de pas af om iets zinnigs over het model te kunnen zeggen. De tweede voorwaarde heeft te maken met de herhaalbaarheid van de metingen. De axioma's (1) tot (4) worden van toepassing geacht op een persoon die de items van een toets beantwoordt. Om de geldigheid van probabilistische uitspraken te onderzoeken zijn veel waarnemingen nodig, maar die kunnen wegens geheugeneffecten niet allemaal bij dezelfde persoon gedaan worden. We zullen dus de antwoorden van meer personen tegelijkertijd moeten analyseren, doch dit impliceert dat we de geldigheid van de axioma's voor alle personen tegelijkertijd veronderstellen. Deze aanname wordt wel eens aangeduid als de aanname van homogeniteit van de populatie. Het is dus belangrijk bij het toetsen van het model niet alleen met een ja of nee als antwoord te komen, doch eveneens aanwijzingen te

vinden dat het model eventueel geldig is in bepaalde subpopulaties en in andere niet. Op dit aspect wordt nader ingegaan in paragraaf 4.5 over persoonsparameterschattingen en in hoofdstuk 9 over itemonzuiverheid.

4.3.2 Relaties tussen het Raschmodel en het multinomiale model

Om een goed begrip te hebben van de statistische toetsen is het nuttig een zeer algemeen statistisch model te beschouwen, waarvan het Raschmodel een speciaal geval is. Indien bij n personen een toets van k items wordt afgenomen levert elke persoon een bepaald antwoord- patroon op. Bij k items zijn er 2^k mogelijke antwoordpatronen en zonder verlies aan informatie kunnen we de observaties samenvatten in een frequentievector met 2^k elementen, waarbij elk element aangeeft hoe vaak het overeenkomstig antwoordpatroon is geobserveerd. Symbolisch duiden we deze frequentie aan met $np_{\boldsymbol{x}}$, waarbij $p_{\boldsymbol{x}}$ de geobserveerde proportie weergeeft van het antwoordpatroon $\boldsymbol{x}$. $p_{\boldsymbol{x}}$ is dus een realisatie van de toevalsvariabele $P_{\boldsymbol{x}}$. Een statistisch model specificeert voor elk antwoordpatroon de kans dat dit antwoordpatroon optreedt. Zie bijvoorbeeld formule (4.56). Kortheidshalve duiden we deze kans aan met $\pi_{\boldsymbol{x}}$. In formule (4.56) is duidelijk dat deze kans een functie is van de modelparameters ε, μ en σ^2. Duiden we nu op een algemene manier het rijtje parameters aan met de vector φ, dan kunnen we expliciet aangeven dat de theoretische kansen een functie zijn van de modelparameters door te schrijven $\pi_{\boldsymbol{x}}(\varphi)$. Het model wordt dus symbolisch geschreven als een vector van 2^k functies van de modelparameters φ, en de observaties zijn vectoren met 2^k overeen- komstige proporties. Deze vectoren worden geschreven als respectievelijk π en $\boldsymbol{p}$. De aannemelijkheidsfunctie kan dus geschreven worden als

$$L(\varphi;\boldsymbol{p},n) = \frac{n!}{(np_1)!\ldots(np_{2^k})!}\prod_{\boldsymbol{x}}\left[\pi_{\boldsymbol{x}}(\varphi)\right]^{np_{\boldsymbol{x}}}, \tag{4.62}$$

waarbij de breuk in het rechterlid aangeeft op hoeveel verschillende manieren de geobserveerde frequentievector uit n observaties gerealiseerd kan worden. Merk op dat deze grootheid niet afhangt van de parameters, en in de aannemelijkheidsfunctie dus als een constante C behandeld kan worden. Het rechterlid van (4.62) is de kansverdeling van de multinomiale verdeling, waarbij de theoretische kansen een functie zijn van de model- parameters. Deze klasse van verdelingen wordt aangeduid als de geparametriseerde multinomiale verdeling.

De eenvoudigste verdeling uit deze familie ontstaat wanneer de theoretische kansen π zelf de parameters zijn. In dat geval spreekt men kortweg van de multinomiale

verdeling. Merk wel op dat niet alle 2^k parameters vrij kunnen variëren, want hun som moet gelijk zijn aan 1; er zijn dus $2^k - 1$ vrije parameters. Evenzo geldt dat er slechts $2^k - 1$ vrije frequenties zijn, want hun som is gelijk aan n. De logaritme van de aannemelijkheidsfunctie in het multinomiaal model is gegeven door

$$\ln L(\pi; \boldsymbol{p}, n) = \ln C + n \sum_{\boldsymbol{x}} p_{\boldsymbol{x}} \ln \pi_{\boldsymbol{x}}, \tag{4.63}$$

waarin men direct de gedaante van de exponentiële familie herkent, waarbij de proporties $p_{\boldsymbol{x}}$ voldoende steekproefgrootheden zijn. De schattingsvergelijkingen zijn dus gegeven door

$p_{\boldsymbol{x}} = \mathscr{E}(P_{\boldsymbol{x}}) = \pi_{\boldsymbol{x}}$ met als oplossing $\hat{\pi}_{\boldsymbol{x}} = p_{\boldsymbol{x}}$.

In dit multinomiale model worden de observaties dus foutloos voorspeld door het model, een betere voorspelling is niet mogelijk. Daarom wordt dit multinomiale model het verzadigde model genoemd.

Keren we nu terug naar het geparametriseerde multinomiale model waar de theoretische kansen $\pi_{\boldsymbol{x}}$ een functie zijn van de parameters φ. In ons voorbeeld bevat φ $k + 1$ vrije para- meters, en voor $k > 2$ geldt dat $k + 1 < 2^k - 1$. Indien we φ vastleggen, liggen alle 2^k theoretische kansen $\pi_{\boldsymbol{x}}$ vast. In de statistiek drukt men dat als volgt uit. In het multinomiale model is de vector π een rijtje van 2^k getallen dat aan zekere voorwaarden moet voldoen. De verzameling van vectoren die aan deze voorwaarden voldoen wordt de parameterruimte genoemd, en deze verzameling duiden we aan met het symbool Ω. In het multinomiale model geldt dus

$$\Omega = \{(\pi_1,...,\pi_{2^k}) \mid \pi_j \geq 0, (j=1,...,2^k); \textstyle\sum_j \pi_j = 1\}. \tag{4.64}$$

In het geparametriseerde multinomiale model brengen we restricties aan op Ω, door te eisen dat de theoretische kansen welbepaalde functies zijn van de parameters φ, in het voorbeeld gegeven door de functieregel (4.56). Deze beperkte parameterruimte duiden we aan met Ω_φ en de definitie is

$$\Omega_\varphi = \{(\pi_1,...,\pi_{2^k}) \mid \pi_j = \pi_j(\varphi), (j=1,...,2^k); \varepsilon_i > 0, (i=1,...,k); \sigma^2 \geq 0\}. \tag{4.65}$$

Aan de hand van formule (4.56) is gemakkelijk na te gaan dat $\pi_{\boldsymbol{x}} \geq 0$ en dat $\sum_{\boldsymbol{x}} \pi_{\boldsymbol{x}} = 1$. Dus elke vector π die behoort tot Ω_φ behoort eveneens tot Ω, of

$$\Omega_\varphi \subset \Omega. \tag{4.66}$$

Als een tweede voorbeeld beschouwen we de CML-schatting van de itemparameters in het Raschmodel. Voor een willekeurig antwoordpatroon $\boldsymbol{x}$ met score s kunnen we steeds schrijven (zie (4.52)) $P(\boldsymbol{x}) = P(\boldsymbol{x}\,|s)P(s)$, of in een wat compactere notatie

$$\pi_{\boldsymbol{x}} = \pi_{\boldsymbol{x}|s}\,\omega_s, \tag{4.67}$$

waarin $\omega_s = P(s)$. Beschouwen we nu een model waarin de frequentievector van de scores de multinomiale verdeling volgt met parameters ω_s, $(s = 0,\ldots,k)$, en de conditionele kansen gegeven zijn door het rechterlid van (4.40), de conditionele kansen in het Raschmodel, dan zien we dat (4.67) een geparametriseerd multinomiaal model definieert met parametervector $\varphi = (\omega_0,\ldots,\omega_s,\ldots,\omega_k,\varepsilon_1,\ldots,\varepsilon_k)$, waarbij echter niet alle parameters vrij zijn, want één itemparameter kunnen we vrij kiezen, en er moet gelden dat $\Sigma_s\omega_s = 1$. Er zijn dus $2k - 1$ vrije parameters in φ. Glas (1989) heeft aangetoond dat de ML-schatters van de ε-parameters de CML-schatters zijn en dat de schatters van de marginale kansen ω_s gegeven zijn door

$$\hat{\omega}_s = p_s, \quad (s = 0,\ldots,k). \tag{4.68}$$

Door de conditionele aannemelijkheid aan te vullen met een verzadigd model voor de scoreverdeling, construeren we een geparametriseerd multinomiaal model. In de volgende paragrafen wordt de statistische toetsingstheorie behandeld waarbij we vaak een beroep zullen doen op deze multinomiale modellen.

4.3.3 Likelihood-ratio-toetsen

Indien een bepaald niet-verzadigd model juist is, kan men niet verwachten dat bij een eindige dataverzameling het maximum van de aannemelijkheidsfunctie even groot zal zijn als het maximum onder het verzadigde model. Immers, het verzadigde model levert altijd het absolute maximum van de aannemelijkheidsfunctie op, terwijl het beperkte model restricties oplegt aan de multinomiale kansen die in een eindige steekproef niet perfect weerspiegeld hoeven te zijn in de geobserveerde proporties. Er geldt dus altijd

$$\frac{L^*(\varphi;\boldsymbol{p},n)}{L^*(\pi;\boldsymbol{p},n)} \leq 1, \tag{4.69}$$

waarin L^* het maximum van de aannemelijkheidsfunctie aanduidt. Anderzijds verwachten we natuurlijk dat, indien het beperkte model juist is, het maximum van de aannemelijkheidsfunctie niet al te zeer zal afwijken van het absolute maximum. De verhouding aangegeven in het linkerlid van (4.69) moet niet al te zeer afwijken van 1, of haar logaritme moet niet al te ver van 0 afwijken. Meer formeel kunnen we de statistische nulhypothese $H_0 : \pi_{\boldsymbol{x}} \in \Omega_\varphi$ toetsen door de overschrijdingskans van (4.69) te bepalen onder de nulhypothese. Deze toets wordt de likelihood-ratio-toets (LR-toets) genoemd. In de theoretische statistiek wordt aan- getoond dat minus twee maal de logaritme van (4.69), vaak aangeduid als G^2, asymptotisch chi-kwadraat verdeeld is indien de nulhypothese waar is. G^2 is dus gegeven door

$$\begin{aligned} G^2 &= 2[\ln L^*(\pi;\boldsymbol{p},n) - \ln L^*(\varphi;\boldsymbol{p},n)] \\ &= 2n\sum_{\boldsymbol{x}} p_{\boldsymbol{x}} \ln \frac{p_{\boldsymbol{x}}}{\hat{\pi}_{\boldsymbol{x}}}, \end{aligned} \tag{4.70}$$

waarin $\hat{\pi}_{\boldsymbol{x}} = \pi_{\boldsymbol{x}}(\hat{\varphi})$, de functie $\pi_{\boldsymbol{x}}$ geëvalueerd op de ML-schatter van φ. Het aantal vrijheidsgraden is het aantal geschatte parameters in het verzadigde model minus het aantal vrije parameters in het beperkte model. In het geval van MML-schattingen is dit dus $[2^k - 1] - [k + 1] = 2^k - k - 2$; in het geval van CML-schattingen is dit verschil $[2^k - 1] - [2k - 1] = 2^k - 2k$. De uitdrukking dat G^2 asymptotisch chi-kwadraat verdeeld is betekent dat de steekproevenverdeling van G^2 goed door de chi-kwadraatverdeling benaderd wordt als n groot wordt; als n niet zeer groot is kan deze benadering slecht zijn, en het gebruik van de chi-kwadraatverdeling dus onterecht. Het probleem is echter wat er precies bedoeld wordt met groot. Het aantal mogelijke antwoordpatronen stijgt zeer snel met het aantal items. Indien $k = 10$ zijn er meer dan 1000 verschillende antwoordpatronen, doch in het sociaalweten- schappelijk onderzoek in Nederland wordt een steekproef van 1000 personen doorgaans als groot beschouwd. In zo'n situatie zal er meestal een vrij groot aantal antwoordpatronen helemaal niet voorkomen in de steekproef, terwijl voor veel andere antwoordpatronen de geobserveerde frequentie klein zal zijn. Of in zo'n geval de chi-kwadraatverdeling een goede benadering is van de verdeling van G^2 is een vraagstuk waar nog veel discussie over is (zie bijv. Read & Cressie, 1988). De schijnbaar voor de hand liggende oplossing om de steekproef dan maar veel groter te maken, heeft echter naast het kostenaspect nog een ander nadeel. Door de steekproefomvang te laten toenemen vergroot ook het onderscheidend vermogen van de statistische toets, dit is de kans om modelafwijkingen te ontdekken. Nu is het natuurlijk wel zo dat men met het construeren van formele modellen, zoals het Raschmodel, hoopt een acceptabele beschrijving te krijgen van de werkelijkheid

met een beperkt aantal concepten, doch het zou heel naïef zijn te denken dat een eenvoudig model de werkelijkheid tot in de kleinste details correct kan weergeven. Als we nu de steekproef heel groot laten worden, wordt de statistische toets ook gevoelig voor onbelangrijke modelafwijkingen, zodat het model steeds verworpen zal worden. De toetsingsgrootheid G^2 zoals gedefinieerd in (4.70) is dus niet goed bruikbaar in de praktijk.

We kunnen echter de LR-toets uitbreiden tot gevallen waarbij het verzadigd model vervangen wordt door een model dat reeds zekere beperkingen oplegt aan Ω, doch waarin we voldoende vertrouwen hebben. We zullen een toets bespreken die door Andersen (1973a) is ontwikkeld, en die geschikt is voor het geval met CML-schatters gewerkt wordt. In paragraaf 4.2.3 werd er op gewezen dat het grote voordeel van de CML-schattingsmethode erin gelegen is dat geen representatieve steekproef hoeft te worden getrokken. Dit impliceert dat, indien het Raschmodel geldig is in een bepaalde populatie, de parameters geschat kunnen worden uit de antwoorden van een willekeurige steekproef, en dat de schattingen binnen de grenzen van de steekproeffout aan elkaar gelijk moeten zijn. Als nu een gegeven steekproef opgedeeld wordt in $k-1$ substeekproeven, waarin voor elke substeekproef geldt dat iedereen dezelfde score heeft, dan kunnen de itemparameters geschat worden uit de antwoorden van elke substeekproef afzonderlijk. Die schattingen moeten ongeveer gelijk zijn aan elkaar, en aan de schattingen die we verkrijgen door de hele steekproef in één keer te analyseren. Dat 'ongeveer gelijk' kunnen we preciseren door een LR-toets te construeren. Even terzijde dient opgemerkt te worden dat de antwoordpatronen met alle items juist of alle items fout geen informatie over de items bevatten. Deze antwoordpatronen kunnen uit de steekproef ver- wijderd worden.

Als algemeen model nemen we aan dat het Raschmodel geldig is in elke subpopulatie afzonderlijk. Binnen elk van de $k-1$ scoregroepen, voor de scores 1 tot $k-1$, moeten dus $k-1$ vrije itemparameters geschat worden. De parametervector duiden we aan met φ_u - de u staat voor 'unrestricted' - en is gegeven door

$$\begin{aligned}\varphi_u &= (\varepsilon_1^{(1)},\varepsilon_2^{(1)},\ldots,\varepsilon_k^{(1)},\varepsilon_1^{(2)},\ldots,\varepsilon_i^{(s)},\ldots,\varepsilon_k^{(k-1)}) \\ &= (\boldsymbol{\varepsilon}^{(1)},\ldots,\boldsymbol{\varepsilon}^{(k-1)}),\end{aligned} \tag{4.71}$$

waarin $\varepsilon_i^{(s)}$ de parameter is van item i in de scoregroep met score s. In de vector φ_u zijn $k(k-1)$ elementen opgenomen omwille van de symmetrie in de notatie, doch er zijn slechts $(k-1)^2$ vrije parameters. Omdat de $k-1$ scoregroepen onafhankelijk zijn

van elkaar kan de aanneme- lijkheidsfunctie voor alle observaties samen geschreven worden als

$$L(\varphi_u; \boldsymbol{X} \mid \boldsymbol{s}) = \prod_{s=1}^{k-1} L(\boldsymbol{\varepsilon}^{(s)}; \boldsymbol{X}^{(s)} \mid s). \tag{4.72}$$

Indien één enkel lid van de familie van Raschmodellen voor alle scoregroepen geldig is, betekent dit dat de itemparameters voor item *i* in alle scoregroepen aan elkaar gelijk moeten zijn. We voeren dus de restrictie in

$$\boldsymbol{\varepsilon}^{(1)} = \ldots = \boldsymbol{\varepsilon}^{(s)} = \ldots = \boldsymbol{\varepsilon}^{(k-1)} = \boldsymbol{\varepsilon} \tag{4.73}$$

en de parametervector φ_r in het beperkte model, waarbij de *r* staat voor 'restricted', is gegeven door

$$\varphi_r = (\varepsilon_1, \ldots, \varepsilon_k). \tag{4.74}$$

Het is duidelijk dat de parameterruimte in het beperkte model een deelverzameling is van de parameterruimte in het algemene model. De restrictie (4.73) is de statistische nulhypothese. Bovendien is het beperkte model niets anders dan het Raschmodel zoals we het tot nog toe behandeld hebben. De toetsingsgrootheid

$$\begin{aligned} Z &= -2 \ln \frac{L^*(\varphi_r; \boldsymbol{X} \mid \boldsymbol{s})}{L^*(\varphi_u; \boldsymbol{X} \mid \boldsymbol{s})} \\ &= 2 \left[\sum_{i=1}^{k-1} \ln L^*(\boldsymbol{\varepsilon}^{(i)}; \boldsymbol{X}^{(i)} \mid s = i) - \ln L^*(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s}) \right] \end{aligned} \tag{4.75}$$

is asymptotisch chi-kwadraat verdeeld met als aantal vrijheidsgraden het verschil in aantal vrije parameters in φ_u min het aantal vrije parameters in φ_r, dus $(k-1)^2 - (k-1) = (k-1)(k-2)$. Indien de waarde van *Z* klein is, betekent dit dat het maximum van de aannemelijk- heidsfunctie niet belangrijk afneemt indien de restrictie (4.73) wordt ingevoerd; men zou kunnen zeggen dat de gegevens zich niet tegen deze restrictie verzetten, en dat we ze dus redelijkerwijze kunnen aannemen.

Om de toetsingsgrootheid *Z* uit te rekenen, moeten de parameters dus *k* keer geschat worden: één keer in elke scoregroep afzonderlijk en één keer voor alle scoregroepen samen. Indien in één van de scoregroepen de parameters niet schatbaar zijn, bijvoorbeeld omdat een item door niemand of door iedereen juist beantwoord is, kan de toetsingsgrootheid niet berekend worden. Om dit probleem op te lossen kan

men ook een LR-toets construeren door verschillende scoregroepen samen te nemen. Stel dat er G scoregroepen gevormd worden, dan veronderstelt het algemene model dat het Raschmodel geldig is in elke der G score- groepen afzonderlijk. De vector φ_u bevat dus $G(k-1)$ vrije parameters. De toetsingsgrootheid wordt uitgerekend op dezelfde manier als in (4.75) is aangegeven, met dien verstande dat de som in het rechterlid G termen bevat. Het aantal vrijheidsgraden is $(G-1)(k-1)$. Andersen (1973a) toont aan dat de toets gevoelig is voor schendingen van axioma (4), dit wil zeggen dat de toets ernaar zal tenderen een significant resultaat op te leveren als de items niet gelijkelijk discrimineren. Indien men scoregroepen samenneemt is het aan te bevelen aan- liggende scoregroepen in dezelfde groep op te nemen. Van den Wollenberg (1982) heeft laten zien dat de toets niet erg gevoelig is voor schendingen van de unidimensionaliteit.

In principe kan men ook een LR-toets construeren indien men met MML-schatters werkt, in plaats van met CML. Het uitrekenen van de toetsingsgrootheid is echter niet eenvoudig met de bestaande programmatuur. Immers het algemene model heeft als parametervector

$$\varphi_u = (\boldsymbol{\varepsilon}^{(0)},...,\boldsymbol{\varepsilon}^{(k)},\mu,\sigma^2),$$

we veronderstellen wel verschillende itemparameters in de verschillende scoregroepen, doch we nemen tevens aan dat de θ-waarden van alle personen in de steekproef een aselecte trekking zijn uit één enkele normale verdeling. De veronderstelling dat er met elke scoregroep een normale verdeling geassocieerd is, doet erg geforceerd aan. Dit betekent dat φ_u uit alle data samen geschat moet worden en daar is de bestaande programmatuur niet op gebouwd. Praktisch gezien is de LR-toets dus beperkt tot het geval dat er CML-schatters voorhanden zijn.

Uit statistisch oogpunt is er geen dwingende reden om de totale steekproef op te delen in homogene scoregroepen. De opdeling kan ook gebeuren volgens een extern criterium, bijvoorbeeld het geslacht of de leeftijd van de respondenten. Voor het gebruik van de LR-toets in zo'n geval verwijzen we naar Andersen (1980).

Een tweede toets, die door Martin-Löf (1973) is ontwikkeld, is wel gevoelig voor schending van het axioma van unidimensionaliteit. Om de toets onderscheidingsvermogen te geven moet men echter een goede hypothese hebben over welke items de verschillende dimensies vertegenwoordigen. Stel dat een toets bestaande uit k items, k_1 kale sommen bevat en k_2 redactiesommen, en dat men vermoedt dat de vaardigheid om de kale sommen op te lossen toch iets anders voorstelt dan de vaardigheid om de redactiesommen op te lossen. Een willekeurig antwoordpatroon $\boldsymbol{x}$ kunnen we schrijven

als $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$, waarbij $\boldsymbol{x}^{(1)}$ het partiële antwoordpatroon is op de k_1 kale sommen en $\boldsymbol{x}^{(2)}$ het partiële antwoordpatroon op de k_2 redactiesommen. Het algemene model, geformuleerd als een geparametriseerd multinomiaal model geeft als kans voor een antwoordpatroon $\boldsymbol{x}$ met s_1 juiste antwoorden in $\boldsymbol{x}^{(1)}$ en s_2 juiste antwoorden in $\boldsymbol{x}^{(2)}$

$$\pi_{\boldsymbol{x}} = P(\boldsymbol{x}^{(1)} \mid s_1)\, P(\boldsymbol{x}^{(2)} \mid s_2)\, \omega_{s_1 s_2},$$

waarin $\omega_{s_1 s_2}$ de kans is op een antwoordpatroon met subscores s_1 respectievelijk s_2. In totaal moeten dus $(k_1 - 1) + (k_2 - 1) = k - 2$ vrije itemparameters geschat worden en $(k_1 + 1)(k_2 + 1) - 1 = k_1 k_2 + k$ vrije multinomiale parameters. De schattingen voor de itemparameters zijn de CML-schattingen die men verkrijgt door de twee subtoetsen met k_1 respectievelijk k_2 items afzonderlijk te analyseren. De schatters van de multinomiale parameters zijn gegeven door

$$\hat{\omega}_{s_1 s_2} = \frac{n_{s_1 s_2}}{n}.$$

Het beperkte model is niets anders dan het Raschmodel, aangevuld met een verzadigd multinomiaal model voor de scoreverdeling, berekend op beide toetsen samen. Dit model heeft $k - 1$ vrije itemparameters en k vrije multinomiale parameters, samen dus $2k - 1$. Het verschil in aantal vrije parameters tussen algemeen en beperkt model is dus $k_1 k_2 - 1$, en dat is ook het aantal vrijheidsgraden voor de toetsingsgrootheid

$$A = 2\left[\sum_{s_1=0}^{k_1}\sum_{s_2=0}^{k_2} n_{s_1 s_2} \ln(n_{s_1 s_2}/n) + \ln L^*(\boldsymbol{\varepsilon}^{(1)}; \boldsymbol{X}^{(1)} \mid \boldsymbol{s}_1) + \ln L^*(\boldsymbol{\varepsilon}^{(2)}; \boldsymbol{X}^{(2)} \mid \boldsymbol{s}_2) - \sum_{s=0}^{k} n_s \ln(n_s/n) - \ln L^*(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s})\right]. \tag{4.76}$$

Merk op dat in formule (4.76) de superscripten wijzen op een opdeling van de items in twee deeltoetsen, terwijl in (4.75) de superscripten wijzen op een opdeling van de steekproef van personen in deelgroepen.

4.3.4 Wald-toetsen

Bij de likelihood-ratio-toetsen hebben we gezien dat het maximum van de aannemelijkheids- functie onder het beperkte model niet al te veel kleiner mag zijn dan het

maximum onder het algemene model om het beperkte model aanvaardbaar te maken. Bij de Wald-toetsen gaat men uit van de volgende rationale: stel dat het beperkte model zegt dat twee parameters β_i en β_j aan elkaar gelijk moeten zijn, doch men schat de parameters zonder die gelijkheid op te leggen, dan mag men verwachten dat de schattingen van die twee parameters niet veel van elkaar zullen verschillen, indien het beperkte model waar is. Men verwacht eigenlijk dat het verschil tussen die twee schattingen uitsluitend veroorzaakt is door de steekproeffout. De nulhypothese luidt dus

$$H_0: \beta_i - \beta_j = 0.$$

Het linkerlid van deze gelijkheid is een functie van de parameters, en de nulhypothese stelt dat deze functie gelijk is aan 0. Nu kunnen we deze nulhypothese complexer maken door niet één functie te beschouwen, maar q functies tegelijkertijd waarbij q niet groter mag zijn dan het aantal vrije parameters. We beschouwen een concreet voorbeeld, dat verder in hoofdstuk 11 wordt besproken. Stel dat een onderzoeker twee Raschtoetsen van k items wil construeren die sterk parallel zijn. Daartoe trekt hij uit een grote itembank k paren van items, zodat binnen elk paar de itemparameters gelijk zijn. Om nog eens te controleren of er werkelijk aan de eis van sterke parallelliteit is voldaan, voegt hij alle items samen in één toets van $2k$ items. Neem aan dat de paren gevormd worden door de items i en $k+i$ $(i = 1, \ldots, k)$. De nulhypothese van de onderzoeker luidt dus

$$H_0: \begin{cases} h_1(\beta) = \beta_1 - \beta_{k+1} = 0 \\ \quad . \\ \quad . \\ h_i(\beta) = \beta_i - \beta_{k+i} = 0 \\ \quad . \\ \quad . \\ h_k(\beta) = \beta_k - \beta_{2k} = 0\,. \end{cases} \tag{4.77}$$

Er geldt dus $q = k$, en het aantal vrije parameters is $2k - 1$. Deze q functies kunnen we verzamelen in een q-vector $\boldsymbol{h}(\beta)$ en de nulhypothese luidt dus in deze compacte notatie:

$$H_0: \boldsymbol{h}(\beta) = \mathbf{0}. \tag{4.78}$$

Beschouw nu de toetsingsgrootheid

$$W = \boldsymbol{h}'(\hat{\beta})[T'(\hat{\beta})\hat{\Sigma}T(\hat{\beta})]^{-1}\boldsymbol{h}(\hat{\beta}), \tag{4.79}$$

waarin T een $2k \times q$ matrix is met elementen t_{ij} gedefinieerd door

$$t_{ij} = \frac{\partial h_j(\beta)}{\partial \beta_i}. \tag{4.80}$$

Σ is de variantie-covariantiematrix van de parameterschatters, en het dakje duidt aan dat alle functies geëvalueerd moeten worden op het punt van de ML-schatters. Wald (1943) heeft aangetoond dat W asymptotisch chi-kwadraat verdeeld is met q vrijheidsgraden, als de nul- hypothese waar is. In het algemeen is het aantal vrijheidsgraden gelijk aan het aantal lineair onafhankelijke restricties die samen de nulhypothese vormen. Het uitrekenen van deze toetsingsgrootheid is niet erg moeilijk omdat de geschatte covariantiematrix meestal voor- handen is als resultaat van de schattingsprocedure. Uit (4.77) volgt direct dat

$$\frac{\partial h_j(\beta)}{\partial \beta_i} = \begin{cases} 1 \text{ indien } i = j, \\ -1 \text{ indien } i = j+k, \\ 0 \text{ in andere gevallen}. \end{cases} \tag{4.81}$$

De matrix T' kan dus geschreven worden als de supermatrix $[I_k | -I_k]$, en de matrix $T'\Sigma T$ is gegeven door

$$T'\Sigma T = \begin{bmatrix} I_k & -I_k \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_k \\ -I_k \end{bmatrix}$$

$$= \Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}.$$

Bij een significant resultaat is het heel natuurlijk om te gaan onderzoeken of het gebrek aan parallelliteit niet te wijten is aan één of meer specifieke itemparen. Dit kan men doen door de k gelijkheden in (4.77) achtereenvolgens als nulhypothese te hanteren en te toetsen. Voor elke afzonderlijke toets geldt dus dat $q = 1$, en de matrix T is een $2k \times 1$ matrix. De matrix $T'\Sigma T$ is dus een 1×1 matrix, en de toetsingsgrootheid W_j krijgt de eenvoudige vorm

$$W_j = \frac{(\hat{\beta}_j - \hat{\beta}_{j+k})^2}{\mathrm{var}(\hat{\beta}_j) + \mathrm{var}(\hat{\beta}_{j+k}) - 2\,\mathrm{cov}(\hat{\beta}_j, \hat{\beta}_{j+k})}, \quad (j = 1, \ldots, q), \tag{4.82}$$

waarin var(.) en cov(.,.) respectievelijk de variantie en covariantie aanduiden. W_j is asymptotisch chi-kwadraat verdeeld met 1 vrijheidsgraad, en $\pm\sqrt{W_j}$ is dus asymptotisch standaardnormaal verdeeld. Het teken ± beduidt dat de vierkantswortel hetzelfde algebraïsch teken krijgt als het verschil $\hat{\beta}_j - \hat{\beta}_{j+k}$ in de teller van (4.82).

Men zou natuurlijk ook kunnen starten met het uitvoeren van de k één-vrijheidsgraad toetsen, en de berekening van de meer ingewikkelde toetsingsgrootheid W achterwege laten. Dit kan men doen als men de volgende overwegingen in acht neemt: de toetsingsgrootheden W_j zijn niet onafhankelijk van elkaar. Hun som is niet gelijk aan W, en de som is ook niet chi-kwadraat verdeeld. Maar de toetsingsgrootheden W_j zijn ook niet volledig afhankelijk van elkaar. Dit betekent dat, indien alle q nulhypothesen waar zijn, de kans dat minstens één toets significant zal uitvallen groter is dan het nominaal significantieniveau α. Men kan dan bijvoorbeeld de Bonferroni toetstechniek gaan gebruiken waar bij de q afzonderlijke toetsen een significantieniveau van α/q wordt gehanteerd, doch deze techniek leidt meestal tot een zeer conservatieve globale toets: de kans dat een fout van de eerste soort gemaakt wordt is weliswaar niet groter dan α, maar kan heel veel kleiner zijn, met als gevolg dat het onderscheidingsvermogen van de toets onnodig klein is. Een toetsingsprocedure die uitgewerkt is door Hommel (1983), neemt dit onnodig strenge criterium weg, terwijl de kans op een fout van de eerste soort toch niet groter is dan α. Voor elk van de q toetsingsgrootheden W_j kan de overschrijdingskans p_j worden uitgerekend. Deze overschrijdingskansen worden geordend van klein naar groot. Deze geordende overschrijdingskansen worden aangeduid als $p_{(j)}$. Dus $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$. De algemene nulhypothese (4.77) wordt verworpen indien

$$p_{(j)} \leq \frac{j\alpha}{q C_q}, \tag{4.83}$$

waarin $C_q = \sum_{j=1}^{q} \frac{1}{j}$.

Tabel 4.5 bevat een voorbeeld, waarbij $q = 5$. $C_5 = 2.283$ en α wordt op 0.05 gesteld.

Tabel 4.5
Voorbeeld van Hommels toetsingsprocedure

j	W_j	p_j	$p_{(j)}$	$(j\alpha)/(qC_q)$
1	0.748	.387	.008	.0044
2	4.019	.045	.017	.0088
3	7.033	.008	.045	.0131
4	1.840	.175	.175	.0175
5	5.696	.017	.387	.0219

Hoewel van drie toetsingsgrootheden W_j de overschrijdingskans kleiner is dan α, leidt de procedure niet tot verwerping van de nulhypothese (4.77) op niveau α. Natuurlijk is het ook mogelijk dat men a priori verdenking koestert tegen de hypothese van parallelliteit van één of meer specifieke paren van items. In zo'n geval is het wel zinvol deze specifieke hypothesen te toetsen op het nominale α-niveau van 5%.

Het is wellicht interessant even na te gaan dat de hypothese (4.77) ook nog op een andere manier getoetst kan worden. Men had bijvoorbeeld de twee deeltoetsen aan twee onafhankelijke steekproeven kunnen aanbieden. In de schattingsprocedure worden de parameters van beide steekproeven dan afzonderlijk geschat. Noemen we de covariantiematrices van de schatters in beide steekproeven Σ_{11} respectievelijk Σ_{22}, dan volgt uit het feit dat de twee steekproeven onafhankelijk zijn van elkaar dat de matrix Σ in (4.79) gegeven is door

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{bmatrix},$$

de submatrices Σ_{12} en Σ_{21} zijn nul-matrices. Voor de toetsingsgrootheden W_j is de covariantieterm in de noemer dus ook gelijk aan 0, waardoor we bij onafhankelijke steekproeven krijgen dat

$$W_j = \frac{(\hat{\beta}_j - \hat{\beta}_{j+k})^2}{\text{var}(\hat{\beta}_j) + \text{var}(\hat{\beta}_{j+k})}. \qquad (4.84)$$

Let wel: de items in de tweede steekproef zijn genummerd $k+1,\ldots,2k$. Hoewel beide toetsingsgrootheden (4.82) en (4.84) allebei asymptotisch chi-kwadraat verdeeld zijn met

1 vrijheidsgraad, zijn beide toetsingsprocedures niet equivalent. Indien de nulhypothese niet waar is, heeft de toetsingsprocedure met afhankelijke steekproeven een veel groter onder-scheidend vermogen dan de procedure met onafhankelijke steekproeven. De toetsings- procedure met onafhankelijke steekproeven heeft echter interessante toepassingen bij het onderzoek naar itemonzuiverheid. Deze toepassingen worden besproken in hoofdstuk 9.

Een toetsingsgrootheid die erg lijkt op W_j zoals gedefinieerd in (4.84) is voorgesteld door Fischer en Scheiblechner (1970), en wordt soms aangeduid als de Fischer-Scheiblechner z_i-toetsingsgrootheid. Hoewel deze toetsingsgrootheid dezelfde formele gedaante heeft als de vierkantswortel-met-teken van (4.84) is er toch een belangrijk verschil. De varianties in de noemer van (4.84) dienen berekend te worden uit de inverse van de informatiematrix. Fischer en Scheiblechner gebruiken echter alleen de hoofddiagonaal van de informatiematrix, dit is, ze gebruiken het kwadraat van (4.59) om de variantie uit te rekenen. Als de schattingen in beide steekproeven gecentreerd worden, dan wordt hierdoor de variantie waarschijnlijk overschat, en is hun toetsings-grootheid dus te klein. Zie voor een exact resultaat bij items van gelijke moeilijkheid paragraaf 4.2.5 en vooral tabel 4.4.

De nulhypothese (4.77) kan ook getoetst worden met een likelihood-ratio-toets. Immers (4.77) is een restrictie op de parameterruimte en de parameters kunnen geschat worden zon-
der en met deze restrictie. Zonder in te gaan op de technische details van het schatten onder restricties, zie daarvoor hoofdstuk 5, is het duidelijk dat voor het construeren van de LR-toets twee maal geschat moet worden, terwijl voor de Wald-toetsen alleen onder het algemene model geschat hoeft te worden. Indien we bovendien de afzonderlijke hypothesen $h_j = 0$ (j=1,...,k) zouden willen toetsen met een LR-toets, dan moeten voor elke hypothese de parameters met die specifieke restrictie opnieuw worden geschat. Voor de toetsing van de k afzonderlijke hypothesen moeten dus $k+1$ schattingsproce-dures uitgevoerd worden, terwijl de Wald-toetsen slechts één enkele schatting vereisen, wat een belangrijke werkbesparing betekent. Bovendien is er een zeer interessant resultaat uit de theoretische statistiek, dat zegt dat beide toetsen asymptotisch equivalent zijn. Dit betekent dat als n toeneemt, de toetsings- grootheden voor beide toetsen ongeveer dezelfde waarde zullen aannemen. De vrijheidsgraden voor beide toetsen zijn gelijk: het aantal restricties q in de Wald-toetsen is precies gelijk aan het verschil in het aantal vrije parameters tussen het algemene model en het beperkte model. Hoewel de keuze tussen de twee procedures voor de hand lijkt te liggen, is het opmerkelijk dat in de bestaande programmatuur bijna geen mogelijkheden zijn voorzien om de Wald toetsen routinematig uit te voeren.

4.3.5 Veralgemeende Pearson X^2-toetsen

De uitkomst van likelihood-ratio-toetsen en van Wald-toetsen is van de data afhankelijk. Bij de likelihood-ratio-toetsen worden de maxima van de aannemelijkheidsfunctie gebruikt onder verschillende restricties op de parameters, maar deze maxima zelf zijn afhankelijk van de data. Bij de Wald-toetsen wordt een functie $\boldsymbol{h}$ berekend op de schattingen van de parameters, en deze schattingen zijn eveneens van de data afhankelijk. Het verband tussen de toetsingsgrootheid en de data is in beide toetsen echter niet zeer doorzichtig. Bij de toetsen die in deze paragraaf worden besproken is het verband tussen de toetsingsgrootheid en de data veel duidelijker: de predicties die uit het model volgen worden op een directe manier met de data vergeleken. De toetsen zijn een veralgemening van de welbekende chi-kwadraat-toetsen die gebruikt worden bij de analyse van contingentietabellen. Allereerst wordt ingegaan op de algemene theorie van deze toetsen. Daarna wordt de theorie op verschillende wijzen toegepast op het Raschmodel, en dit levert toetsen op die gevoelig zijn voor bepaalde schendingen van het Raschmodel.

Algemene theorie

Hoewel de chi-kwadraat-toetsen in de sociale wetenschappen routinematig worden toegepast, kan het nuttig zijn even in te gaan op de theorie achter die toetsen. Daarom beginnen we met een voorbeeld. Stel dat we willen nagaan of de antwoorden op twee vragen in een enquête statistisch afhankelijk zijn van elkaar. De observaties waarover we beschikken zijn weergegeven in tabel 4.6. De eerste variabele kan drie waarden aannemen, *a*, *b* en *c*; de tweede variabele kan de waarden *A* en *B* aannemen. De eerste variabele duiden we aan met *X*, en de uitspraak $X=a$ betekent dus dat de eerste variabele de waarde *a* aanneemt. De tweede variabele zullen we aanduiden met *Y*. In het corpus van de tabel staan bivariate frequenties: voor 25 personen uit de steekproef geldt de uitspraak "$X=a$ en $Y=B$".

Tabel 4.6
Tweedimensionale contingentietabel

a	*b*	*c*	totaal

A	25	17	2	44
B	67	42	9	118
totaal	92	59	11	162

We kunnen van de tweedimensionale tabel 4.6 gemakkelijk een ééndimensionale tabel maken door de frequenties achter elkaar te schrijven. Dit is gebeurd in tabel 4.7.

Tabel 4.7

Tweedimensionale tabel omgevormd tot een ééndimensionale tabel

aA	*bA*	*cA*	*aB*	*bB*	*cB*
25	17	2	67	42	9

Door dit te doen, definiëren we impliciet een nieuwe variabele Z die zes verschillende waarden kan aannemen, zoals aangeduid in de bovenste regel van tabel 4.7. Het spreekt vanzelf dat beide tabellen precies dezelfde informatie bevatten. De uitspraak "$Z=aB$" is dus equivalent met de gecombineerde uitspraak over de twee oorspronkelijke variabelen "$X=a$ en $Y=B$", de waarden van Z zijn dus antwoordpatronen, en tabel 4.7 bevat de geobserveerde frequenties van alle zes mogelijke antwoordpatronen.

Om te onderzoeken of de variabelen X en Y afhankelijk zijn van elkaar, moeten we zorgvuldig een aantal stappen zetten. We moeten een model formuleren, de parameters van het model schatten, een toetsingsgrootheid definiëren en nagaan wat de overschrijdingskans is van de uit de gegevens berekende toetsingsgrootheid. Het eenvoudigste, verzadigde model is dat de zes frequenties uit tabel 4.6 een multinomiale verdeling volgen: bij een aselecte trekking uit de populatie is er de kans $\pi_{ij} = P(X=i, Y=j)$, $(i = a,b,c; j = A,B)$ dat de observatie in cel (i,j) van tabel 4.6. terechtkomt. Omdat de som van de kansen gelijk moet zijn aan 1, betekent dit dat in het verzadigde model vijf parameters geschat moeten worden. De ML-schatters in het multinomiale model zijn gelijk aan de celproporties: $\hat{\pi}_{ij} = n_{ij}/n$, zodat onmiddellijk duidelijk is dat het model de geobserveerde frequenties perfect voorspelt. Om de afhankelijkheid te onderzoeken, stellen we een nulhypothese op die afhankelijkheid ontkent. De variabelen X en Y zijn stochastisch onafhankelijk indien:

$$\pi_{ij} = \pi_i \pi_j, \quad (i = a,b,c; j = A,B) \tag{4.85}$$

waarin $\pi_i = P(X=i)$ en $\pi_j = P(Y=j)$. Omdat $\Sigma_i \pi_i = \Sigma_j \pi_j = 1$, zijn er in het beperkte model slechts drie parameters. Hun ML-schatters zijn gegeven door de relatieve frequenties

van de marginale totalen: $\hat{\pi}_i = n_i/n$ en $\hat{\pi}_j = n_j/n$. In het beperkte model is de ML-schatter van π_{ij} dan gegeven door:

$$\hat{\pi}_{ij} = \hat{\pi}_i \hat{\pi}_j = \frac{n_i n_j}{n^2} \tag{4.86}$$

en de verwachte frequentie in de *(i,j)*-de cel van tabel 4.6 is gegeven door de welbekende formule:

$$E_{ij} = n\hat{\pi}_{ij} = \frac{n_i n_j}{n}. \tag{4.87}$$

Indien de restrictie (4.85) geldig is, mogen de verwachte frequenties E_{ij} niet al te veel afwijken van de geobserveerde frequenties O_{ij}, niet meer dan door de steekproeffout kan worden verklaard. Pearson heeft aangetoond dat de toetsingsgrootheid

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.88}$$

asymptotisch chi-kwadraat verdeeld is. Het aantal vrijheidsgraden is gelijk aan het aantal vrije cellen in de tabel verminderd met het aantal geschatte parameters. In het voorbeeld dus 5-3=2. De grootheid X^2, berekend op de gegevens van tabel 4.6, bedraagt 0.53, terwijl de kritieke waarde voor α=0.05 in de chi-kwadraatverdeling met twee vrijheidsgraden 5.99 is. Er is dus geen reden om het model van onafhankelijkheid (4.85) te verwerpen. Het is belangrijk het aantal termen in de som van het rechterlid van (4.88) niet te verwarren met het aantal vrije cellen. Er moet gesommeerd worden over alle cellen van de tabel en niet alleen over de vrije cellen.

Er is vrij uitvoerig op dit voorbeeld ingegaan, opdat duidelijk zou worden dat er een aantal stappen is gezet die in de routinematige uitvoering van de toets vaak niet meer worden opgemerkt. We becommentariëren deze stappen een voor een.

(1) Er is steeds sprake van een model, en van restricties op de parameterruimte. Pearson heeft zijn toets ontwikkeld voor het geval het model een multinomiaal model is. Daarom is het belangrijk bij toepassingen van Pearsons toets steeds precies na te gaan of het model waarmee men werkt beschouwd kan worden als een multinomiaal model. De nulhypothese komt steeds overeen met een restrictie op de parameterruimte. In het voorbeeld is deze restrictie gegeven door (4.85). Het is belangrijk op te merken dat Pearsons toets niet beperkt is tot deze restrictie alleen. De methode die Pearson heeft ontworpen is geldig voor een zeer grote klasse van restricties. Voor alle gevallen die in dit boek worden beschouwd,

kan de methode worden toegepast. Een uiteenzetting van de statistische theorie kan men vinden in hoofdstuk 14 van Bishop, Fienberg en Holland (1975). Men zou bijvoorbeeld het beperkte model (4.85) nog verder kunnen beperken met de extra eis:

$$\pi_a = \pi_b = \pi_c = 1/3. \tag{4.89}$$

(2) Er moeten parameters geschat worden, en deze parameters worden geschat onder de nulhypothese. Gebruiken we bijvoorbeeld (4.85) en (4.89) samen als nulhypothese, dan hoeft alleen de parameter π_A te worden geschat, want de andere parameters zijn precies vastgelegd. Merk bovendien op dat de parameters geschat worden uit dezelfde data als waarop de grootheid X^2 wordt berekend.

(3) De verwachte frequenties moeten worden uitgerekend met de schattingen van de para- meters onder de nulhypothese. De eerste gelijkheid in (4.87) is dus algemeen geldig, de tweede gelijkheid niet: deze geldt alleen onder de nulhypothese van onafhankelijkheid. Nemen we (4.85) en (4.89) samen als nulhypothese, dan krijgen we als verwachte frequenties

$$E_{ij} = n\hat{\pi}_{ij} = n\pi_i\hat{\pi}_j = \frac{n_j}{3}. \tag{4.90}$$

(4) De steekproevenverdeling van X^2 in (4.88) is niet bekend. Pearson heeft aangetoond dat, indien n toeneemt deze steekproevenverdeling steeds beter gaat lijken op de theoretische chi-kwadraatverdeling. De chi-kwadraatverdeling wordt dus gebruikt als een benadering voor de echte steekproevenverdeling van X^2. Hoe goed die benadering in concrete gevallen is, weten we niet exact. Wel is door veel onderzoek bekend dat voor praktische doeleinden het gebruik van de chi-kwadraatverdeling gerechtvaardigd is indien n niet al te klein is en indien er niet al te veel cellen zijn met kleine verwachte frequenties. Soms wordt de vuistregel gehanteerd dat het aantal cellen met verwachte frequentie kleiner dan 5 niet meer mag bedragen dan 20% van het aantal cellen. Wat men in zulke gevallen meestal doet is overgaan tot het samennemen van cellen. In tabel 4.6 zou men bijvoorbeeld alle cellen 'b' en 'c' kunnen samennemen, zodat er een 2×2 tabel ontstaat. Deze procedure is zeker gerechtvaardigd, mits men goed in het oog houdt dat hierdoor een nieuwe variabele X' gecreëerd wordt, die niet drie maar slechts twee antwoord- categorieën heeft. Het toepassen van Pearsons toets gebeurt dan op de twee variabelen X' en Y, die samen maar vier waarden kunnen aannemen. Kortom, er wordt een nieuw model geformuleerd, de

parameters worden opnieuw geschat en het besluit dat men trekt is alleen van toepassing op de variabelen X' en Y, en niet op X en Y.

(5) Het besluit dat men neemt, aanvaarden of verwerpen van de nulhypothese, betreft de nulhypothese als geheel. Is de nulhypothese bijvoorbeeld de combinatie van (4.85) en (4.89), die in het voorbeeld zeker verworpen moet worden, dan volgt uit de toetsing niet of de significantie te wijten is aan (4.85) of aan (4.89). Werkt men met heel complexe nulhypothesen, zoals het Raschmodel, dan geeft de toetsingsgrootheid dus niet de mogelijkheid een modelschending precies te lokaliseren. Pearsons toets is dus een globale toets van het model.

Passen we nu het voorgaande toe op het Raschmodel, dan is het vrij eenvoudig om de toetsingsgrootheid X^2 te construeren. Naar analogie met de tabellen 4.6 en 4.7 kunnen we de observaties onderbrengen in een k-dimensionale frequentietabel, of in een unidimensionale tabel. De tweede voorstelling is voor onze doeleinden het handigst. Bij een toets met k items zijn er 2^k antwoordpatronen mogelijk, en elke persoon die de toets beantwoordt, levert precies één antwoordpatroon op. Bij n personen kunnen we dus de frequentie $n_{\boldsymbol{x}}$ bepalen waarmee antwoordpatroon $\boldsymbol{x}$ is opgetreden. Alle frequenties samen volgen dus de multinomiale verdeling; het model is zeker niet verzadigd want er zijn $2^k - 1$ vrije cellen en er zijn maar $k-1$, in het geval van CML, of $k+1$, in het geval van MML, parameters geschat. De grootheid X^2 is dus gegeven door:

$$\begin{aligned} X^2 &= \sum_{\boldsymbol{x}} \frac{(n_{\boldsymbol{x}} - n\hat{\pi}_{\boldsymbol{x}})^2}{n\hat{\pi}_{\boldsymbol{x}}} \\ &= n \sum_{\boldsymbol{x}} \frac{(p_{\boldsymbol{x}} - \hat{\pi}_{\boldsymbol{x}})^2}{\hat{\pi}_{\boldsymbol{x}}}, \end{aligned} \tag{4.91}$$

waarin $p_{\boldsymbol{x}} = n_{\boldsymbol{x}}/n$. X^2 is asymptotisch chi-kwadraat verdeeld met $2^k-1-(k-1) = 2^k - k$ vrij- heidsgraden (CML) of 2^k-k-2 vrijheidsgraden (MML). Het bezwaar tegen het gebruik van deze toetsingsgrootheid is natuurlijk dat reeds bij middelgrote k, zeg 20, het aantal cellen van de tabel vele malen groter zal zijn dan de steekproef, zodat automatisch zeer veel, zo niet alle cellen een heel kleine verwachte waarde zullen hebben. Bij k=20 en n=1000 is de gemiddelde verwachte frequentie kleiner dan .001. Het is wel zeker dat het gebruiken van de chi-kwadraatverdeling als benadering van de verdeling van X^2 niet terecht is. Er zit dus niet veel anders op dan onze toevlucht te nemen tot het samenvoegen van cellen. Doch dan zouden strikt genomen de parameters opnieuw geschat moeten worden, waarbij in de schattings-procedure geen gebruik

gemaakt mag worden van de afzonderlijke frequenties van de samengevoegde cellen. Zo'n schattingsprocedure opzetten is echter vrij moeilijk en omslachtig.

Glas en Verhelst (1989) hebben een methode ontwikkeld om een soort correctie op de gewone grootheid X^2 aan te brengen, zonder dat de parameters opnieuw geschat moeten worden. Bovendien is hun methode algemener toepasbaar dan in de situatie waar cellen worden samengenomen. Bij het samennemen van cellen worden de cellen van de oorspronkelijke contingentietabel ingedeeld in een aantal groepen, en elke van de oorspronkelijke cellen wordt aan precies één groep toegewezen. Bij de methode van Glas en Verhelst is het ook mogelijk bepaalde cellen aan meer groepen groep toe te wijzen of cellen buiten beschouwing te laten. Later zullen we zien dat deze mogelijkheid ons in staat stelt om gerichte toetsen te construeren in plaats van alleen maar een globale toets.

De methode is vrij complex en zal in een aantal stappen worden uiteengezet. Eerst wordt aangetoond hoe Pearsons grootheid X^2 als een matrix-expressie kan worden geschreven. Deze matrix-expressie wordt een kwadratische vorm genoemd. Vervolgens wordt getoond hoe het samennemen of groeperen van cellen kan gebeuren door gebruik te maken van een speciaal daartoe geconstrueerde matrix Y. De toetsingsgrootheid Q, waarmee we gaan werken, is ook een kwadratische vorm. De waarde die deze kwadratische vorm aanneemt is afhankelijk van de observaties, maar ook van de matrix Y die we geconstrueerd hebben. Om deze afhankelijkheid expliciet aan te geven zullen we de toetsingsgrootheid aanduiden als $Q(Y)$. De centrale vraag is natuurlijk of $Q(Y)$ asymptotisch chi-kwadraat verdeeld is, en wat het geassocieerde aantal vrijheidsgraden is. Met een voorbeeld zullen we aantonen dat $Q(Y)$ niet chi-kwadraat verdeeld is voor elke matrix Y. Glas en Verhelst hebben een klasse van Y-matrices gekarakteriseerd waarvoor $Q(Y)$ wel asymptotisch chi-kwadraat verdeeld is. We zullen dit resultaat niet in zijn algemeenheid bespreken, maar ons beperken tot het geval waar het geparametriseerd multinomiaal model tot de exponentiële familie behoort.

Pearsons X^2 als een kwadratische vorm

Om elegant te kunnen werken is het nuttig (4.91) als een matrix-expressie te schrijven. Definieer $m = 2^k$, m is dus het aantal mogelijke antwoordpatronen. De geobserveerde proporties $p_{\boldsymbol{x}}$ worden verzameld in de vector $\boldsymbol{p}$ en de geschatte kansen $\hat{\pi}_{\boldsymbol{x}}$ in de vector $\hat{\pi}$. Bovendien definiëren we een diagonaalmatrix $D_{\hat{\pi}}$, met de elementen van $\hat{\pi}$ op de diagonaal. Het is gemakkelijk na te gaan dat (4.91) geschreven kan worden als:

$$X^2 = n(\boldsymbol{p} - \hat{\pi})' D_{\hat{\pi}}^{-1} (\boldsymbol{p} - \hat{\pi})$$
$$= n(\boldsymbol{p} - \hat{\pi})' I_m (I_m D_{\hat{\pi}} I_m)^{-1} I_m (\boldsymbol{p} - \hat{\pi}), \quad (4.92)$$

waarbij I_m de $m \times m$ identiteitsmatrix is. De algemene gedaante van (4.92) is het produkt van een rijvector met een symmetrische matrix met een kolomvector, waarbij de twee vectoren in het produkt gelijk zijn aan elkaar. Een dergelijk produkt wordt in de lineaire algebra een kwadratische vorm genoemd. Door het toevoegen van de identiteitsmatrix wordt expliciet aangegeven dat de som in (4.91) uit m termen bestaat: elke afwijking tussen geobserveerde (p) en verwachte (π) proportie wordt gekwadrateerd, en draagt dus bij tot de som X^2.

Het samennemen van cellen

De manier waarop cellen moeten worden samengenomen kan worden aangegeven in een speciaal daartoe geconstrueerde matrix Y. De matrix Y in tabel 4.8 is een voorbeeld voor een geval met $k = 3$ items. De matrix bevat alleen enen en nullen, en voorlopig kunnen we er vanuit gaan dat de enen op willekeurige plaatsen zijn neergezet. De acht mogelijke antwoordpatronen zijn afgebeeld onder het kopje T_1; de matrix T_2 komt later aan de orde.
Beschouw nu het produkt $(\boldsymbol{p} - \hat{\pi})' \boldsymbol{y}_2$, waarin $\boldsymbol{y}_2$ de tweede kolom van Y is. Dit produkt geeft de som van de afwijkingen $p_x - \pi_x$ voor het vijfde en het zevende antwoordpatroon, dit is voor de twee antwoordpatronen waarvoor een 1 staat in de overeenkomstige rij van de tweede kolom van Y. Op analoge manier is het produkt $(\boldsymbol{p} - \hat{\pi})' \boldsymbol{y}_1$ de som (met één term) van alle antwoordpatronen waarbij een 1 staat in de eerste kolom van Y. Men kan ook zeggen dat in elke kolom alle afwijkingen meedoen: ze worden eerst vermenigvuldigd met een constante die in hun rij staat. In het voorbeeld zijn die constanten 1 of 0, maar we hadden ook andere constanten kunnen invullen. Het vermenigvuldigen van een aantal elementen, de afwijkingen, met een constante en die produkten bij elkaar optellen geeft een som die men een lineaire combinatie van die elementen noemt. De constanten waarmee vermenigvuldigd is, worden de coëfficiënten genoemd. Het produkt $(p - \hat{\pi})' \boldsymbol{Y}$ definieert dus in het algemeen evenveel
lineaire combinaties als er kolommen zijn in Y. Merk op dat de antwoordpatronen 1, 2, 4,

Tabel 4.8

Constructie van de matrix voor de veralgemeende Pearson toetsen

T_1	T_2	Y

0	0	0	1	0	0	0	0	0
1	0	0	0	1	0	0	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	0	0	0
1	1	0	0	0	1	0	0	1
1	0	1	0	0	1	0	0	0
0	1	1	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0

6 en 8 in geen van beide groepen zijn opgenomen. Het zal duidelijk zijn dat een matrix Y die de antwoordpatronen groepeert in de gebruikelijke zin van het woord, aan de volgende eis moet voldoen: in elke rij van de matrix moet precies één 1 voorkomen, de andere elementen van de rij zijn gelijk aan nul. Het groeperen is dus ook het nemen van een aantal lineaire combinaties.

Beschouw nu de kwadratische vorm

$$Q(Y) = n(\boldsymbol{p} - \hat{\pi})' Y (Y' D_{\hat{\pi}} Y)^{-} Y' (\boldsymbol{p} - \hat{\pi}), \tag{4.93}$$

waarin de aanduiding '-' in superscript een veralgemeende inverse aanduidt. Indien de matrix Y niet van volle rang is, dat wil zeggen, indien één of meer van zijn kolommen kunnen worden geschreven als een lineaire combinatie van de andere kolommen, dan is de matrix $Y' D_{\hat{\pi}} Y$ singulier en heeft geen reguliere inverse. Singuliere matrices hebben echter wel oneindig veel zogenaamde veralgemeende inversen. De kwadratische vorm $Q(Y)$ heeft echter altijd dezelfde waarde, ongeacht welke veralgemeende inverse men kiest. Indien de matrix van de kwadratische vorm niet singulier is, is de inverse matrix uniek. Een vergelijking van (4.93) met (4.92) leert ons onmiddellijk dat $X^2 = Q(I_m)$, dus X^2 is een speciaal geval van (4.93) met $Y = I_m$. Daaruit volgt echter niet dat $Q(Y)$ asymptotisch chi-kwadraat verdeeld is voor elke Y.

$Q(Y)$ is niet voor elke Y chi-kwadraat verdeeld

De antwoordpatronen waarbij een 1 staat in de tweede kolom van de matrix Y in tabel 4.8 kunnen als volgt worden omschreven: het zijn alle antwoordpatronen die een juist antwoord hebben op item 2 en een score 2. Indien de parameters met CML geschat zijn geldt: $\hat{\pi}' \boldsymbol{y}_2 = n^{-1}(n_2 \hat{\pi}_{2|2})$. Voor de geobserveerde proporties geldt analoog dat $\boldsymbol{p}' \boldsymbol{y}_2 = n^{-1}(n_2 p_{2|2})$. De ene 1 in de eerste kolom heeft betrekking op het antwoordpatroon met score 1 en een juist antwoord op item 2, zodat ook hier soortgelijke

uitdrukkingen gelden voor de produkten $\hat{\pi}'\boldsymbol{y_1}$ en $\boldsymbol{p}'\boldsymbol{y}_1$. Omdat in de rijen van de matrix Y nooit meer dan één element verschilt van 0 is de matrix $Y'D_{\hat{\pi}}Y$ een diagonaalmatrix. De kwadratische vorm (4.93) kan dan ook expliciet geschreven worden als

$$Q(Y) = \sum_{s=1}^{2} \frac{n_s(p_{2|s} - \hat{\pi}_{2|s})^2}{\hat{\pi}_{2|s}}. \tag{4.94}$$

Hoewel deze uitdrukking erg lijkt op het laatste lid van (4.91), zijn er enkele belangrijke verschillen. Deze kunnen we het beste toelichten door de score × itemantwoord-contingentietabel te construeren (zie tabel 4.9).

Tabel 4.9

Verwachte frequenties in de score × itemantwoord-tabel voor item 2

	x_2=0	x_2=1
s=0	---	---
s=1	$n_1(1-\hat{\pi}_{2\|1})$	$n_1\hat{\pi}_{2\|1}$
s=2	$n_2(1-\hat{\pi}_{2\|2})$	$n_2\hat{\pi}_{2\|2}$
s=3	---	---

Er zijn twee opmerkelijke verschillen met de situatie die leidde tot formule (4.91). Het eerste is dat in de som (4.94) maar twee termen zijn opgenomen en niet vier, zoals door tabel 4.9 wordt gesuggereerd. Bovendien zijn vier van de mogelijke cellen helemaal uit de kwadratische vorm weggelaten. Nu is het wel zo dat in die vier cellen de score 0 of 3 bedraagt, waardoor de geobserveerde en verwachte frequenties precies aan elkaar gelijk zijn, maar in het algemeen kan natuurlijk een matrix Y geconstrueerd worden waarbij cellen worden weggelaten, waarvoor de overeenkomst tussen geobserveerde en verwachte proporties niet perfect is. De wel ingevulde cellen waarvoor $x_2 = 0$ zijn ten onrechte niet meegeteld.

Het tweede verschil heeft te maken met de parameterschattingen en het aantal vrijheidsgraden. In totaal zijn er vijf vrije parameters geschat: twee itemparameters en drie parameters ω_s voor het verzadigde multinomiale model van de scorefrequenties. In tabel 4.9 zijn vier vrije cellen, en het mechanisch toepassen van de regel voor het bepalen van de vrijheidsgraden zou 4-5=-1 vrijheidsgraden opleveren, hetgeen natuurlijk

onzin is. De vijf parameters kunnen natuurlijk niet geschat worden als alleen de frequenties gegeven zijn die overeenkomen met de ingevulde cellen van tabel 4.9. Dit toont duidelijk aan dat $Q(Y)$ niet asymptotisch chi-kwadraat verdeeld is voor elke willekeurige matrix Y.

Een klasse van Y-matrices waarvoor $Q(Y)$ asymptotisch chi-kwadraat verdeeld is

Glas en Verhelst (1989) hebben een klasse van Y-matrices gekarakteriseerd waarvoor geldt dat $Q(Y)$ asymptotisch chi-kwadraat verdeeld is. Hier geven we alleen het resultaat voor exponentiële-familiemodellen. Om de uiteenzetting niet nodeloos abstract te maken, zullen we de principes eerst uiteenzetten aan de hand van een concreet voorbeeld, het Raschmodel, waarbij de parameters met CML geschat worden. Zoals reeds is opgemerkt zijn de CML-schatters in het Raschmodel equivalent met de gewone ML-schatters van de itemparameters, als we het Raschmodel aanvullen met een verzadigd multinomiaal model voor de scoreverdeling.

Het resultaat van Glas en Verhelst is het gemakkelijkst te begrijpen door gebruik te maken van voldoende steekproefgrootheden. Om te laten zien dat het Raschmodel, aangevuld met een verzadigd multinomiaal model voor de verdeling van de scores een lid van de exponentiële familie is, definiëren we $k+1$ zogenaamde indicatorvariabelen $t_j, j=0,...,k$, die de waarde 1 of 0 kunnen aannemen. De variabele $t_j=1$ indien de score op de k items gelijk is aan j, anders is t_j gelijk aan 0. Merk op dat de waarde van t_j eenduidig uit de antwoord- vector $\boldsymbol{x}$ kan worden berekend. Voorbeeld: als $k=3$ en $\boldsymbol{x}=(1\ 0\ 1)$, dan is de score 2, en de indicatorvector heeft de waarde $t=(0\ 0\ 1\ 0)$. We kunnen dus evengoed zeggen dat de observatie bestaat uit het antwoordpatroon $\boldsymbol{x}$, als uit de combinatie van antwoordpatroon en indicatorvector $(\boldsymbol{x},\boldsymbol{t})$. De uitdrukking (4.67) kunnen we dus ook schrijven als $\pi_{\boldsymbol{x},\boldsymbol{t}} = \pi_{\boldsymbol{x}|\boldsymbol{t}}\,\pi_{\boldsymbol{t}}$, waarin de eerste factor in het rechterlid de conditionele kans op het antwoordpatroon is, gegeven de indicator van de score. De log-aannemelijkheidsfunctie is gegeven door

$$\ln\ L(\boldsymbol{\varepsilon},\pi;\boldsymbol{x},\boldsymbol{t}) = \sum_i x_i \ln \varepsilon_i + \sum_j t_j \ln \omega_j - \ln \gamma_s(\boldsymbol{\varepsilon}) \qquad (4.95)$$

waaruit duidelijk blijkt dat de vector $(\boldsymbol{x},\boldsymbol{t})$ een voldoende steekproefgrootheid is voor de parameters: de vector $\boldsymbol{t}$ is voldoende voor de multinomiale parameters ω_s en de vector $\boldsymbol{x}$ is voldoende voor de itemparameters. Het feit dat de vector $(\boldsymbol{x},\boldsymbol{t})$ $2k+1$ elementen bevat, terwijl er maar $2k$-1 vrije parameters zijn is voorlopig niet belangrijk; we komen er later op terug.

Om er voor te zorgen dat de kwadratische vorm $Q(Y)$ asymptotisch chi-kwadraat verdeeld is, kan aangetoond worden dat de voldoende steekproefgrootheden ($\boldsymbol{x},\boldsymbol{t}$) op een of andere manier te vinden moeten zijn in elke rij van de matrix Y. Dit is, kort samengevat, het belangrijkste resultaat van Glas en Verhelst. Voor de matrix Y in tabel 4.8 is dit zeker niet het geval. Een eenvoudige manier om de voldoende steekproefgrootheden in de matrix te brengen, bestaat erin een gegeven matrix Y uit te breiden met die steekproefgrootheden. Dit is gebeurd in tabel 4.8. De rijen van de matrix T_1 zijn de antwoordpatronen $\boldsymbol{x}$ en de rijen van T_2 zijn de erbij behorende indicatorvectoren $\boldsymbol{t}$. Definieer nu $T=[T_1|T_2]$, en $Z=[T_1|T_2|Y]= [T|Y]$. In plaats van $Q(Y)$ wordt $Q(Z)$ uitgerekend, en omdat in de rijen van Z de afdoende steekproefgrootheden aanwezig zijn, geldt het volgende resultaat:

(1) $Q(Z)=Q([T|Y])$ is asymptotisch chi-kwadraat verdeeld waarbij het aantal vrijheidsgraden gelijk is aan de rang van de matrix Z min 1, min het aantal geschatte parameters. Dit geldt voor elke matrix Y.

Men zou natuurlijk kunnen opperen dat dit allemaal goed en wel is, doch dat daarmee het oorspronkelijke probleem is veranderd. Bij de behandeling van het voorbeeld zijn we immers begonnen met het beschouwen van slechts twee lineaire combinaties van afwijkingen, namelijk $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'\boldsymbol{y}_1$ en $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'\boldsymbol{y}_2$, terwijl de matrix Z negen kolommen heeft, en het produkt $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'Z$ dus negen lineaire combinaties definieert. Er kan echter bewezen worden (Glas, 1989) dat, indien de parameters zijn geschat met de ML methode, geldt:

(2) $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'T = \boldsymbol{0}$. Daaruit volgt onmiddellijk dat $Q(T) = 0$.

De lineaire combinaties die we toegevoegd hebben zijn dus gelijk aan 0. Dit betekent echter niet dat $Q(Y)=Q(Z)$. Het belangrijkste verschil is dat de matrix $Z'D_{\hat{\pi}}Z$ gebruikt moet worden in de kwadratische vorm en niet de diagonale matrix $Y'D_{\hat{\pi}}Y$. De reden hiervoor is dat de parameters uit de oorspronkelijke data geschat zijn en niet uit de lineaire combinaties $\boldsymbol{p}'Y$ die minder informatie bevatten.

Hiervoor werd gezegd dat de voldoende steekproefgrootheden 'aanwezig' moesten zijn in de matrix Z van lineaire combinaties. We hebben ons van die aanwezigheid verzekerd door een gegeven matrix uit te breiden. Dit is een handige methode, maar ze is niet noodzakelijk. De precieze definitie van aanwezig zijn is als volgt. Stel dat een geparametriseerd multinomiaal model met s vrije parameters tot de exponentiële familie behoort. Het aantal verschillende antwoordpatronen is m. Beschouw de $m \times s$

matrix U, waarvan elke rij de minimaal voldoende steekproefgrootheden voor het desbetreffende antwoordpatroon bevat. Voor een gegeven $m \times r$ matrix Z, waarbij $r>s+1$, is de kwadratische vorm $Q(Z)$, gedefi- nieerd door (4.93) asymptotisch chi-kwadraat verdeeld als aan de volgende twee voorwaarden is voldaan:

(3) elke kolom van de matrix U kan geschreven worden als een lineaire combinatie van de kolommen van Z;
(4) de m-vector **1**, dit is de vector waarvan alle elementen gelijk zijn aan 1, kan geschreven worden als een lineaire combinatie van de kolommen van Z.

Voor de matrix $Z=[T_1|T_2|Y]$ uit tabel 4.8 is dit het geval. Er zijn slechts twee vrije itemparameters en drie vrije marginale kansen ω_s. De matrix U kunnen we dus vormen door in de matrix $T=[T_1|T_2]$ bijvoorbeeld de eerste kolom van T_1 en de eerste kolom van T_2 te schrappen. Aan voorwaarde (3) is dan op een triviale manier voldaan. Door de kolommen van de matrix T_2 bij elkaar op te tellen zien we ook dat aan voorwaarde (4) is voldaan.

We beschikken dus over twee manieren om aan te tonen dat, binnen de exponentiële familie, de kwadratische vorm $Q(Z)$ asymptotisch chi-kwadraat verdeeld is: ofwel we breiden een gegeven matrix Y uit met een matrix die de voldoende steekproefgrootheden en de vector **1** bevat, ofwel we tonen aan dat aan de voorwaarden (3) en (4) is voldaan.

Voor een gedetailleerde uiteenzetting van bovenstaande resultaten, zie Glas (1989), Glas en Verhelst (1989) en Verhelst en Eggen (1989).

Praktische problemen

Het resultaat dat hierboven is gegeven, heeft zeer veel toepassingsmogelijkheden omdat de matrix Y die in resultaat (1) staat volkomen willekeurig is. Alle toetsen van het Raschmodel die hierna nog besproken zullen worden, zijn speciale gevallen van (4.93). De algemeenheid van het resultaat dient echter niet overschat te worden, want er duiken een viertal praktische problemen op waarmee men in de toepassing terdege rekening moet houden.

Het eerste probleem heeft te maken met het uitrekenen van de kwadratische vorm $Q(Y)$. De matrix Y heeft $m=2^k$ rijen. Indien we de kwadratische vorm $Q(Y)$ uitrekenen met de matrixvermenigvuldigingen als aangegeven in (4.93), moet gigantisch veel rekenwerk worden uitgevoerd, zelfs voor niet al te grote k. We zullen dus moeten

zoeken naar een aangepaste definitie van de matrix Y waardoor het rekenwerk snel en efficiënt kan verlopen.

Het tweede probleem heeft te maken met het aantal vrijheidsgraden. Dat aantal is gegeven door rang(Y)-s-1, waarin s het aantal vrije parameters van het model is. Het bepalen van de rang van Y moet met de nodige zorgvuldigheid gebeuren. Ook als we de methode van de toegevoegde matrix T gebruiken, en de kwadratische vorm $Q([T|Y])$ beschouwen, is het niet automatisch zo dat het aantal vrijheidsgraden gelijk is aan het aantal kolommen van Y. In het voorbeeld van tabel 4.8 is het aantal vrije parameters s gelijk aan 5, de rang van de matrix $T=[T_1|T_2]$ is s+1=6, maar de rang van $Z=[T|Y]$ is niet 6+2=8, maar 7, omdat de kolom- men van Y lineair afhankelijk zijn van de kolommen van T. Dit kan men in tabel 4.8 gemakkelijk controleren: de som van de twee kolommen van Y is gelijk aan de tweede kolom van T_1 min de laatste kolom van T_2. Het aantal vrijheidsgraden geassocieerd met $Q(Z)$ is dus niet 2 maar 1.

Het derde probleem heeft te maken met het feit dat van $Q(Y)$ alleen de asymptotische verdeling bekend is, maar niet de exacte verdeling. De chi-kwadraatverdeling wordt dus gebruikt als een benadering van de exacte verdeling. Het is echter niet bekend hoe goed die benadering is in concrete gevallen. Het enige wat we eigenlijk kunnen doen, is waarschuwen tegen het gebruik van (4.93) en de chi-kwadraatverdeling bij zeer kleine steekproeven, en het vermijden van lineaire combinaties in de matrix Y die zeer kleine proporties van het totale aantal observaties vertegenwoordigen. Zo is de eerste kolom van de matrix Y in tabel 4.8 een lineaire combinatie waarin alleen het antwoordpatroon (0 1 0) is betrokken. Als het aantal personen in de steekproef met dit antwoordpatroon zeer klein is, kan betwijfeld worden of de chi-kwadraatverdeling wel een goede benadering is van de exacte verdeling van de kwadratische vorm.

Het vierde probleem is het belangrijkste en luidt: "hoe moet men de matrix Y kiezen ?" Het feit dat $Q(Y)$ voor een grote klasse van Y-matrices asymptotisch chi-kwadraat verdeeld is, betekent niet dat het er niet toe doet welke matrix we uit die klasse kiezen. De kwadratische vorm is alleen chi-kwadraat verdeeld onder de nulhypothese, dat wil zeggen indien het model waar is. Indien één of meer veronderstellingen van het model geschonden zijn, is het onderscheidend vermogen van de statistische toets afhankelijk van de matrix Y die we gekozen hebben. Dit kunnen we reeds zien door een inspectie van formule (4.94). De afwijkingen die in de teller staan hebben betrekking op item 2. Het is dus te verwachten dat het gebruik van de matrix Y uit tabel 4.8 een toets zal opleveren die vooral gevoelig is indien er, in termen van het model, iets mis is met item 2, eerder dan met item 1 of item 3.

Bij de specifieke toetsen voor het Raschmodel die hierna worden besproken, zal aan deze vier problemen aandacht worden geschonken.

De S_i-toetsen

De S_i-toetsen zijn bedoeld om modelschendingen op itemniveau te kunnen ontdekken. Voor elk item wordt een toets geconstrueerd, en de matrix Y heeft betrekking op een bepaald item. In deze paragraaf wordt dit specifieke item aangeduid met de index i. Om dit expliciet aan te geven krijgt de matrix Y een index i mee. Deze toetsen zijn alleen van toepassing indien de parameters met de CML-methode zijn geschat.

Het totale scorebereik wordt opgedeeld in r intervallen, dat wil zeggen de scores worden opgedeeld in r scoregroepen van aaneengesloten scores. Daarbij mogen de score 0 en de perfecte score buiten beschouwing gelaten worden. Deze scoregroepen duiden we aan als de verzamelingen G_q, q=1,...,r. Bijvoorbeeld, stel k=10 en r=3, dan is een mogelijke opdeling G_1={1,2,3,4}, G_2={5,6} en G_3={7,8,9}. De matrix Y heeft r kolommen waarbij elke kolom overeenkomt met een scoregroep. De waarden in de Y_i-matrix zijn 0 of 1; een 1 in de q-de kolom wordt ingevuld voor elke rij (antwoordpatroon) indien de score van dit antwoordpatroon behoort tot de q-de scoregroep, en indien het een antwoordpatroon betreft met een juist antwoord op item i. De matrix Y in tabel 4.8 is volgens deze regel geconstrueerd, waarbij r=2, G_1={1}, G_2={2} en i=2. Merk op dat uit deze regel volgt dat in elke rij van de Y-matrix niet meer dan één element kan verschillen van 0. Dit heeft het prettige voordeel dat de matrix $Y_i' D_{\hat{\pi}} Y_i$ een diagonale matrix is. De kolommen van Y_i zijn echter lineair afhankelijk van de kolommen van T, zoals hierboven reeds is aangetoond. Definiëren we nu twee vectoren met lineaire combinaties van afwijkingen tussen $\boldsymbol{p}$ en π:

$$\boldsymbol{d}_1 = (\boldsymbol{p} - \hat{\pi})' T, \quad \boldsymbol{d}_2 = (\boldsymbol{p} - \hat{\pi})' Y_i,$$

dan weten we uit de vorige paragraaf dat $\boldsymbol{d}_1 = \boldsymbol{0}$. Door een vrij lange afleiding, die we hier niet bespreken, zie Verhelst en Eggen (1989) voor details, kan aangetoond worden dat de kwadratische vorm $Q([T|Y_i])$ gegeven is door:

$$Q([T|Y_i]) = n\boldsymbol{d}'_2 [Y_i' D_{\hat{\pi}} Y_i - \Delta_i - A_i]^{-} \boldsymbol{d}_2. \tag{4.96}$$

De matrix Δ_i in (4.96) is een $r \times r$ diagonale matrix waarvan de elementen op de diagonaal gegeven zijn door

$$(\Delta_i)_{qq} = \sum_{s \in G_q} \frac{n_s}{n} \hat{\pi}^2_{i|s}. \tag{4.97}$$

De matrix A_i is een symmetrische $r \times r$ matrix waarvan de elementen afhankelijk zijn van de informatiematrix, zie (4.48). De precieze definitie van de elementen van A_i is nogal omslachtig en wordt hier achterwege gelaten. Theoretisch gezien echter is deze matrix uiterst belangrijk, omdat hij precies de correctie bevat die noodzakelijk is, omdat de parameters niet zijn geschat uit de gegevens die bevat zijn in een score × itemantwoord-contingentietabel, maar uit de oorspronkelijke data, die meer informatie bevatten. Bovendien is het uitrekenen van de matrix A_i in de praktijk een tijdrovend karwei, dat bij grote k zelfs niet goed meer uit te voeren is. Daarom stellen we ons vaak tevreden met een benaderende kwadratische vorm door de matrix A_i in (4.96) gewoon weg te laten. Deze benaderende kwadratische vorm kan geschreven worden als:

$$Q^*([T|Y_i]) = n\boldsymbol{d}_2'[Y_i' D_{\hat{\pi}} Y_i - \Delta_i]^{-1}\boldsymbol{d}_2$$

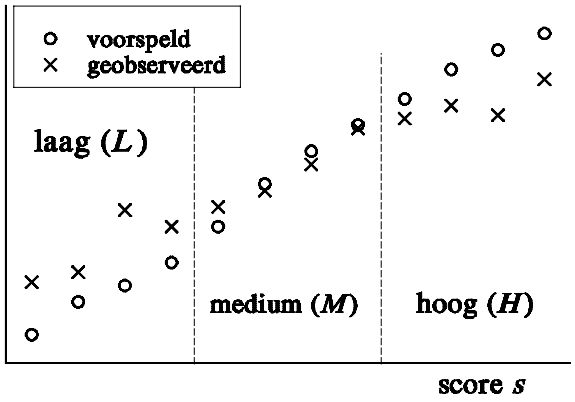
$$= \sum_{q=1}^{r} \frac{\left[\sum_{s \in G_q} n_s(p_{i|s} - \hat{\pi}_{i|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})} . \qquad (4.98)$$

De kwadratische vorm $Q([T|Y_i])$ is asymptotisch chi-kwadraat verdeeld met r-1 vrijheids- graden; van de benaderende vorm Q^* gegeven in (4.98) is de asymptotische verdeling niet bekend. Ervaring heeft echter geleerd dat beide grootheden heel vaak niet veel van elkaar afwijken, maar dat de vorm Q^* meestal een iets grotere uitkomst oplevert. Door Q^* te inter- preteren als een chi-kwadraat verdeelde variabele met r-1 vrijheidsgraden zal men dus de nulhypothese iets vaker verwerpen dan aangegeven door het nominale significantieniveau α.

In het vervolg zullen we de kwadratische vorm $Q([T|Y_i])$ aanduiden als S_i en de benaderende grootheid $Q^*([T|Y_i])$ als S_i^*.

Een nadere beschouwing van de teller in het rechterlid van (4.98) kan ons iets leren over het onderscheidend vermogen van de S_i-toetsen. De uitdrukking tussen vierkante haken is een som van afwijkingen tussen geobserveerde en verwachte frequenties. Deze afwijkingen kun- nen positief of negatief zijn. Indien nu binnen een scoregroep G_q zowel positieve als nega- tieve afwijkingen voorkomen, dan heffen die elkaar (ten dele) op. Doordat alleen hun som wordt gekwadrateerd is het dus mogelijk dat grote afwijkingen door dit compensatiemechanisme slechts een geringe bijdrage leveren aan de toetsingsgrootheid. Of er compensatie optreedt, is afhankelijk van de manier van groeperen in scoregroepen. In figuur 4.7 is een voorbeeld gegeven van een item dat slechter discrimineert dan door het Raschmodel is voorspeld.

De geobserveerde proporties, gezien als functie van de score, vertonen een vlakker verloop dan de voorspelde proporties. De verticale stippellijnen in de figuur geven aan dat er drie scoregroepen zijn, die zijn aangeduid als laag, medium en hoog. Omdat de modelafwijkingen systematisch zijn, zien we dat in de twee extreme groepen geen compensatie optreedt, de afwijkingen hebben allemaal hetzelfde teken; in de medium-groep echter zal de som van de afwijkingen nagenoeg nul zijn. Deze groep draagt dus weinig of niets bij aan de toetsingsgrootheid S_i. Hadden we de twee extreme groepen, laag en hoog, als één enkele groep behandeld, door de twee overeenkomstige kolommen in de matrix Y_i bij elkaar op te tellen, dan zou in deze gecombineerde groep ook cancellatie optreden, en de resulterende kwadratische vorm zou nauwelijks van nul verschillen.



Figuur 4.7
Een item dat slechter discrimineert dan voorspeld door het Raschmodel

Aan dit voorbeeld zien we dat het onderscheidend vermogen van de toets afhankelijk is van de manier waarop de scoregroepen gevormd worden en de bijbehorende Y matrix wordt geconstrueerd. Men zou nu kunnen denken dat maximaal onderscheidend vermogen bereikt kan worden door eerst een plaatje te construeren analoog aan figuur 4.7, en dan de groepsindeling te maken zodanig dat er geen cancellatie van positieve en negatieve afwijkingen optreedt binnen de scoregroepen. Of andersom, als men liever geen significantie heeft, de groepen zo maken dat er zoveel mogelijk cancellatie optreedt. Op zo'n manier echter wordt de toetsingsprocedure afhankelijk gemaakt van de data, of preciezer gezegd, van de afwijkingen tussen geobserveerde en voorspelde frequenties. Dus is de Y-matrix geen matrix van constanten maar een matrix van toevalsvariabelen waarvan de waarde van steekproef tot steekproef zal gaan verschillen. Maar in dat geval is de toetsingsgrootheid S_i niet meer chi-kwadraat verdeeld. In de

praktijk echter zal men er niet helemaal onderuit kunnen om de groepsindeling toch enigszins van de data te laten afhangen. De noemer van het rechterlid van (4.97) zal klein zijn indien voor alle scores in G_q de geobserveerde frequenties zeer klein zijn of de verwachte proporties $\hat{\pi}_{i/s}$ zeer dicht bij 0 of 1 liggen. Het is twijfelachtig of in zo'n geval de benadering door de chi-kwadraatverdeling nog wel gerechtvaardigd is. Door een andere groepsindeling te kiezen kan men die kleine noemers vermijden. Maar een groepsindeling 'op maat' vereist dat de data geconsulteerd worden. Hoewel een dergelijke handelwijze niet helemaal orthodox is, maakt ze de S_i-toetsen niet waardeloos. Immers om de groepsindeling zo te maken dat de noemer van (4.97) niet al te klein wordt, hoeven de afwijkingen tussen geobserveerde en verwachte proporties niet geconsulteerd te worden. In het programma OPLM (Verhelst, Glas & Verstralen, 1993) wordt de minimale waarde van de noemers in (4.97) op 5 gesteld.

In de literatuur zijn verschillende toetsingsgrootheden voorgesteld waarvan de formule erg veel lijkt op het rechterlid van (4.98). We noemen als voorbeelden Wright en Panchapakesan (1969), Bock (1972), Wright en Mead (1977), Elliott, Murray en Saunders (1977) en Yen (1981). Er zijn echter twee belangrijke punten waarop de toetsingsgrootheden van al deze auteurs verschillen van (4.98).

Het eerste is de wijze waarop de verwachte proporties worden uitgerekend. Wij gebruiken de conditionele kans gegeven de score, en deze kans is onafhankelijk van θ; bovengenoemde auteurs gebruiken echter allemaal een schatting die gebaseerd is op een schatter van θ, die bovendien gebaseerd is op een JML-procedure. Deze benadering heeft het schijnbare voordeel dat de toetsen dan ook gebruikt kunnen worden voor andere modellen dan het Raschmodel, zoals het twee- en drieparameter-logistische model, doch het bewijs dat de toetsingsgrootheden asymptotisch chi-kwadraat verdeeld zijn ontbreekt, en de bewering is waarschijnlijk ook onjuist. In ieder geval kan men voor het bewijs geen beroep doen op standaardresultaten uit de statistiek, want die vereisen allemaal schatters met bepaalde eigenschappen. Een van de eisen is consistentie van de parameterschatters. In het Raschmodel zijn JML-schatters niet consistent en voor het tweeparameter-logistische model is geen bewijs van consistentie gegeven. Afgezien hiervan hebben alle formules die door bovengenoemde auteurs worden gepresenteerd in de teller dezelfde gedaante als het rechterlid van (4.98).

Het tweede punt is dat de noemers nogal verschillen. Wright en Panchapakesan (1969) presenteren dezelfde noemer als in (4.98), doch hun toets is alleen ontworpen voor het Raschmodel waarbij scores niet worden gegroepeerd. De noemer van (4.98) is een som van varianties, waarbij elke term de variantie is van het aantal juiste antwoorden in de scoregroep met s juiste antwoorden. In de toets die Yen (1981) voorstelt, wordt deze som vervangen door de variantie van het aantal juiste items in de

groep, waarbij gedaan wordt alsof alle personen in de groep dezelfde kans op een juist antwoord hebben. Het effect hiervan is dat de noemer te groot wordt. Wright en Mead (1971) houden hier rekening mee, en voeren een correctiefactor in. Hun formule heeft in de noemer dezelfde gedaante als de noemer van (4.98). De meest afwijkende vorm komt voor in de formule die Elliott e.a. (1977) gebruiken: daar bevat de noemer geen varianties maar verwachte aantallen juiste antwoorden. Hun toetsingsgrootheid is te vergelijken met (4.94), en komt erop neer dat in termen van contingentietabellen de helft van de cellen ten onrechte niet meegeteld wordt. Hun toetsingsgrootheid is dan ook systematisch veel te klein. Een overzicht van al deze formules wordt gegeven door Yen (1981).

De M_i-toetsen

Stel dat we een item onderzoeken dat beter discrimineert dan het merendeel van de andere, en we construeren voor dit item een figuur analoog aan figuur 4.7, dan zullen we zien dat de geobserveerde proporties een steiler verloop vertonen dan de verwachte, maar de S_i-toets kan geen onderscheid maken tussen te grote en te kleine discriminatie, want in beide gevallen is de toetsingsgrootheid positief. Er kunnen natuurlijk nog andere afwijkingen optreden die niet zo'n systematisch patroon te zien geven, maar die, als ze voldoende groot zijn, ook een significant (positief) resultaat opleveren. Door een slimme constructie van de matrix Y_i kan onderscheid gemaakt worden tussen items die te weinig en die te veel discriminerend vermogen hebben. De scores worden opgedeeld in drie groepen, een laag-, een medium- en een hoog-groep, precies zoals in figuur 4.7 is aangegeven. De Y_i-matrix bestaat echter uit één enkele kolom, waar een 1 staat indien de score van het antwoordpatroon een juist antwoord bevat op item *i*, en de bijbehorende score tot de laag-groep behoort. In geval de score tot de hoog-groep behoort, vult men -1 in en voor de medium-groep komt overal 0 te staan. De kwadratische vorm $Q([T|Y_i])$ is asymptotisch chi-kwadraat verdeeld met één vrijheidsgraad. De 'vierkantswortel-met-teken', dat wil zeggen, de positieve vierkantswortel vermenigvuldigd met -1 indien de één-elements vector $\boldsymbol{d}_2$ negatief is, volgt dus de standaardnormale verdeling. De benaderende waarde van deze toetsingsgrootheid, gebaseerd op (4.98), is gegeven door

$$M_i^* = \frac{\sum_{s\in L} n_s(p_{i|s} - \hat{\pi}_{i|s}) - \sum_{s\in H} n_s(p_{i|s} - \hat{\pi}_{i|s})}{\left[\sum_{s\in L,H} n_s\hat{\pi}_{i|s}(1-\hat{\pi}_{i|s})\right]^{½}}, \tag{4.99}$$

waarin L en H verwijzen naar respectievelijk de laag- en de hoog-groep. Uit figuur 4.7 volgt duidelijk dat in die situatie de eerste som in de teller van (4.99) een positieve waarde zal aannemen, en de tweede som een negatieve waarde. Het verschil zal dus een positieve waarde krijgen, en omdat de noemer van (4.99) steeds positief is, krijgen we dus bij een te weinig discriminerend item een positieve uitkomst. Bij een te sterk discriminerend item zal de uitkomst negatief zijn.

Door de bovenstaande omschrijving liggen de M-toetsen echter niet eenduidig vast, omdat de begrippen laag-groep en hoog-groep niet nauwkeurig gedefinieerd zijn. In het programma OPLM worden drie varianten van de M-toetsen uitgerekend, waarbij drie verschillende definities van laag-groep en hoog-groep worden gehanteerd. De drie toetsingsgrootheden worden aangeduid als respectievelijk M_i, $M2_i$ en $M3_i$. De definities van de verschillende score groepen is als volgt:

M_i : $s \in$ L indien $\hat{\pi}_{i|s} \leq 0.4$ en $s \in$ H indien $\hat{\pi}_{i|s} \geq 0.6$;

$M2_i$: de scores worden in een laag-groep en een hoog-groep verdeeld zodanig dat $\sum_{s \in L} n_s \approx \sum_{s \in H} n_s \approx n/2$. De medium-groep is leeg. Het is niet steeds mogelijk dat precies de helft van de observaties in beide groepen valt, omdat alle antwoordpatronen met dezelfde score tot dezelfde groep moeten behoren;

$M3_i$: analoog aan de situatie bij $M2_i$, doch nu is de opdeling in drie groepen die elk ongeveer een derde van de observaties vertegenwoordigen.

Door Molenaar (1983) is een toets ontwikkeld die als een speciale variant van de hier besproken M-toetsen kan worden opgevat. In de inleiding van deze paragraaf hebben we gezien dat de matrix Y een willekeurige matrix is. Indien we in een bepaalde rij een 1 invullen, en in een andere rij 2, blijven de theoretische resultaten geldig. Alleen kennen we verschillende gewichten toe aan verschillende antwoordpatronen. Molenaar stelt voor de afwijkingen $n_s(p_{i|s} - \hat{\pi}_{i|s})$ te wegen met het omgekeerde van hun standaardafwijking. Op de plaatsen waar in de Y-vector voor de M_i-toetsen een 1 of -1 komt, plaatst Molenaar de grootheid $\pm[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}$, waarbij de positieve wortel genomen wordt voor de laag-groep en de negatieve voor de hoog-groep. De toetsingsgrootheid, door Molenaar U_i, genoemd is gegeven door

$$U_i = \frac{\sum_{s \in L} \frac{n_s(p_{i|s} - \hat{\pi}_{i|s})}{[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}} - \sum_{s \in H} \frac{n_s(p_{i|s} - \hat{\pi}_{i|s})}{[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}}}{(|L| + |H|)^{\frac{1}{2}}} \qquad (4.100)$$

waarin $|L|$ en $|H|$ het aantal verschillende scores is in respectievelijk de laag- en de hoog-groep. Het is niet moeilijk om aan te tonen dat U_i hetzelfde is als $Q^*(T|Y_i])$, met verschillende gewichten in de een-koloms matrix Y_i. De U_i-toetsen zijn geïmplemen-

teerd in het programma PML (Gustafsson, 1979, aanpassing door Molenaar, 1981). Voor deze U_i-toetsen wordt ook een andere definitie van de laag-groep en de hoog-groep gebruikt dan in de *M*-toetsen. De laag-groep bevat de 25% laagst scorende en de hoog-groep de 25% hoogst scorende observaties.

De R_{1c}-toets

Hoewel de S_i-toetsen allemaal asymptotisch chi-kwadraat verdeeld zijn, zijn ze niet onafhankelijk van elkaar. Dit betekent dat hun som niet chi-kwadraat verdeeld is. Bovendien moet men voorzichtig zijn bij de interpretatie van de S_i-toetsen. Indien het model geldig is, dan kan men verwachten dat ongeveer 100α% van de toetsen een significant resultaat zal opleveren bij toetsen op niveau α. Dit resultaat is niet exact, omdat de toetsen niet onafhankelijk zijn van elkaar. De kans dat een of meer toetsen significant zijn is echter behoorlijk groter dan het nominale significantieniveau α. Om een globale toets te construeren kan men de toetsingsprocedure van Hommel gebruiken die reeds werd besproken in paragraaf 4.3.4, of men kan gebruik maken van een globale toets die beschouwd kan worden als een combinatie van alle S_i-toetsen. Deze toets is de R_{1c}-toets die door Glas (1989) werd ontwikkeld.

De rationale van deze toets is uiterst eenvoudig: hij is niets anders dan de kwadratische vorm $Q(Y)$, gegeven door (4.93), waarbij $Y = [Y_1|Y_2|...|Y_k]$.

Het uitrekenen van deze kwadratische vorm is in het algemeen echter zeer ingewikkeld omdat de matrix $Y' D_{\hat{\pi}} Y$ niet langer diagonaal is. Dit is precies de reden waarom de S_i-toetsen niet onafhankelijk zijn van elkaar. Glas (1989) heeft aangetoond dat een belangrijke vereenvoudiging aangebracht kan worden indien de opdeling in scoregroepen G_q voor alle items dezelfde is. In tabel 4.10 zijn de drie Y_i-matrices afgebeeld voor een toets met drie items, waarbij echter de kolommen gepermuteerd zijn. Elke kolom draagt een dubbele index *iq*, waarbij de eerste index verwijst naar het item en de tweede naar de scoregroep. Er zijn ook maar zes rijen afgebeeld, omdat de antwoordpatronen (0 0 0) en (1 1 1) niets aan de toetsingsgrootheid bijdragen. Indien men de parameters schat met CML komt het weglaten van die antwoordpatronen overeen met het aannemen van een verzadigde multinomiale verdeling van de scorefrequenties voor de scores 1, 2,...,*k*-1. Blokken van de totale *Y*-matrix die volledig uit nullen bestaan zijn wit gelaten.

Het is gemakkelijk na te gaan dat de matrix $Y' D_{\hat{\pi}} Y$ in dit geval een blokdiagonale structuur heeft, waarbij elk blok betrekking heeft op één scoregroep. Bovendien is gemakkelijk in te zien dat de kolommen van de matrices T_1 en T_2 geschreven kunnen

worden als lineaire combinaties van de kolommen van Y. De i-de kolom van de matrix T_1 in tabel 4.8
is gegegeven door $Y_i1 + Y_{i2}$, de tweede kolom van T_2 is gegeven als $\Sigma_i Y_{i1}$ en de derde ko-

Tabel 4.10
De Y-matrix voor de R_{1c}-toets (k=3)

Y_{11}	Y_{21}	Y_{31}	Y_{12}	Y_{22}	Y_{32}
1	0	0			
0	1	0			
0	0	1			
			1	1	0
			1	0	1
			0	1	1

lom als $\Sigma_i Y_{i2}/2$. De eerste en de laatste kolom van T_2 kunnen buiten beschouwing worden gelaten omdat de patronen met score 0 en 3 verwijderd zijn. De matrix Y bevat dus de matrix T, als lineaire combinaties van zijn kolommen, en daarom is $Q(Y)$ asymptotisch chi-kwadraat verdeeld. Het aantal vrijheidsgraden is hier 3, en in het algemeen $k(r-1)$. De benaderende vorm $Q^*(Y)$, in dit geval aangeduid als R_{1c}^*, is een eenvoudige veralgemening van (4.98):

$$R_{1c}^* = Q^*(Y) = \sum_{q=1}^{r} \sum_{i=1}^{k} \frac{\left[\sum_{s \in G_q} n_s (p_{i|s} - \hat{\pi}_{i|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})}. \qquad (4.101)$$

Meestal is de benaderende vorm $Q^*(Y)$ groter dan de exacte vorm $Q(Y)$; de asymptotische verdeling is echter niet bekend. Uit een vergelijking van (4.98) en (4.101) is direct duidelijk dat, indien voor alle items dezelfde groepering is gebruikt, geldt dat

$$R_{1c}^* = \sum_i S_i^* .$$

In de literatuur is op verschillende plaatsen aan deze globale toets aandacht gegeven. Martin-Löf (1973) heeft een zogenaamde T-toets ontwikkeld, vanuit een iets andere rationale dan hier werd gebruikt (zie bijvoorbeeld Van den Wollenberg, 1979). Er kan echter aangetoond worden (Glas, 1981) dat Martin-Löfs T-toets equivalent is met de R_{1c}-toets. De R_{1c}-toets is geïmplementeerd in het programma OPLM, de T-toets wordt uitgerekend in het programma PML. Merk echter op dat beide toetsingsgrootheden, uitgerekend met dezelfde data niet noodzakelijkerwijze dezelfde uitkomst geven: de

uitkomst is natuurlijk afhankelijk van de wijze waarop de scores zijn gegroepeerd, en dit gebeurt in de twee programma's niet op identieke wijze.

Van den Wollenberg (1979, 1982) heeft de Q_1-toets voorgesteld. De toetsingsgrootheid Q_1 is een kleine modificatie van (4.101):

$$Q_1 = \frac{k-1}{k} R_{1c}^{*} \quad .$$

Uit simulatiestudies blijkt dat de verdeling van Q_1 goed te benaderen is door de chi-kwadraat verdeling.

Bij het gebruik van de R_{1c}-toets dient men aan twee zaken aandacht te geven. In de eerste plaats is dat de grootte van de noemer in (4.101). Door het feit dat voor de R_{1c}-toets dezelfde scoregroepering gebruikt wordt voor alle items, is het soms onvermijdelijk dat één of meer noemers in (4.101) zeer klein worden, waardoor sommige termen erg groot worden. In zo'n geval is het twijfelachtig of nog wel een beroep gedaan kan worden op de chi-kwadraat verdeling. Het tweede probleem betreft het gecombineerde gebruik van itemgerichte toetsen, bijvoorbeeld de S_i-toetsen, en een globale toets als R_{1c}. Het is mogelijk dat de R_{1c}-toets niet significant is, terwijl één of meer S_i-toetsen een zeer significant resultaat opleveren. De reden hiervoor is dat de R_{1c}-toets minder onderscheidend vermogen heeft dan de S_i-toetsen voor zeer specifieke modelschendingen. Men zou kunnen stellen dat de R_{1c}-toets een 'slecht' item niet opmerkt als het ingebed is in een toets waarvan de meeste items aan het Raschmodel voldoen. Omgekeerd is het ook mogelijk dat de modelschendingen niet zonder meer aan specifieke items kunnen worden toegeschreven, zodat de itemgerichte toetsen niet significant zijn, maar bijvoorbeeld in meerderheid een kleine overschrijdingskans hebben, bijvoorbeeld kleiner dan 0.5. In zo'n geval kan de 'niet zo schitterende prestatie' van de afzonderlijke S_i- toetsen gecombineerd worden in de R_{1c}-toets die wel tot significantie kan leiden. Daarom is het in de praktijk aan te raden itemgerichte toetsen en globale toetsen gecombineerd te gebruiken.

Van den Wollenberg (1979, 1982) heeft laten zien dat de R_{1c}- (of de Q_1-) toets niet erg geschikt is om schendingen van het unidimensionaliteitsaxioma te ontdekken. Een theoretisch eenvoudige generalisatie van de R_{1c}-toets, namelijk de R_{2c}-toets is wel gevoelig voor deze schendingen. De teller van (4.98) en (4.101) bevat zogenaamde-eerste-orde-afwijkingen $n_s(p_{i|s}-\hat{\pi}_{i|s})$. Nu kan ook een toetsingsgrootheid worden opgesteld die tweede-orde-afwijkingen onderzoekt: de proportie personen die zowel item *i* als item *j* juist beantwoordt, wordt vergeleken met de voorspelde proportie. Er wordt dus een vector $\boldsymbol{d}$ van afwijkingen opgesteld die als elementen de afwijkingen $n_s(p_{ij|s}-\hat{\pi}_{ij|s})$ heeft, voor alle scores s=2,...,k-2 en voor alle paren (i,j), i>j=1,...,k. De bijbehorende Y-matrix heeft dan $rk(k-1)/2$ kolommen, en voor grote k is de R_{2c}-

toetsingsgrootheid niet goed uit te rekenen. Details over de berekeningswijze kan men vinden in Glas (1989). Van den Wollenberg (1979, 1982) geeft een benaderende toetsingsgrootheid Q_2.

De R_0- en de R_{1m}-toetsen

De S_i-toetsen, de M_i-toetsen, en de R_{1c}-toets zijn allemaal toepasbaar indien de parameters geschat zijn met de CML-schattingsmethode. Gebruiken we echter MML, dan ligt de zaak heel wat gecompliceerder. Immers, MML is niet zomaar een methode, maar veronderstelt een ander model dan alleen maar het Raschmodel; er dient een hypothese toegevoegd te worden over de verdeling van de latente variabele θ. De combinatie van het Raschmodel en de verdeling van θ zorgt er voor dat het model als geheel niet meer tot de exponentiële familie behoort, en dat we voor de constructie van statistische toetsen niet zonder meer een beroep kunnen doen op de resultaten (1) en (2) die hiervoor werden gegeven.

Voor de normale verdeling geldt wel resultaat (1), namelijk dat $Q([T|Y])$ asymptotisch chi-kwadraat verdeeld is indien T is opgebouwd volgens de beschrijving die hiervoor werd gegeven. Het tweede resultaat, namelijk $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T = \mathbf{0}$, geldt echter niet meer. Glas (1989) heeft in zijn onderzoekingen geconstateerd dat $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_1 = \mathbf{0}$, zonder dat hij evenwel deze gelijkheid in het algemeen kon bewijzen. Bij gebruik van MML is echter de vector $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_2 \neq \mathbf{0}$. Met behulp van tabel 4.8 is het gemakkelijk na te gaan dat $n\boldsymbol{p}'T_2$ niets anders is dan de $(k+1)$-vector met geobserveerde scorefrequenties $(n_0, n_1, \cdots, n_k)$, dus de vector $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_2$ geeft de afwijkingen aan tussen de geobserveerde en voorspelde proportie van elke score. Bij CML was de overeenkomst perfect door het invoeren van een verzadigd multinomiaal model met k parameters. Door de invoering van de veronderstelling van een normale verdeling van θ zal de overeenkomst niet meer perfect zijn. Als de hypothese van een normale verdeling echter juist is, moeten de afwijkingen toe te schrijven zijn aan de steekproeffout. Dus de grootheid

$$R_0 = Q([T_1|T_2]) \tag{4.102}$$

is asymptotisch chi-kwadraat verdeeld. Het aantal vrijheidsgraden is k-2. De R_0-toets is gevoelig voor schendingen van de normaliteitsassumptie.

De R_{1m}-toets wordt op precies dezelfde manier geconstrueerd als de R_{1c}-toets. De afwijkingen tussen voorspelde en geobserveerde proporties kunnen nu echter toegeschreven worden zowel aan schendingen van het Raschmodel, dus de combinatie van S_i-achtige toetsen, als aan schendingen van de assumptie van normaliteit van de

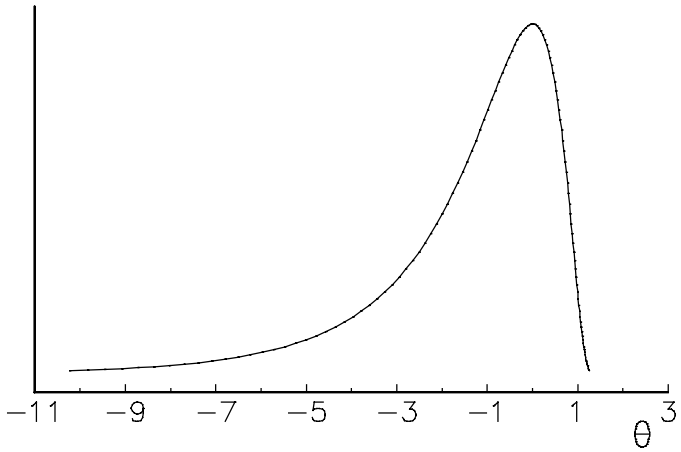
verdeling van theta. Het aantal vrijheidsgraden van R_{1m} bedraagt dan ook k-2 meer dan van de R_{1c}-toets: de k multinomiale parameters ω_s zijn niet meer nodig, doch worden vervangen door de twee parameters van de normale verdeling. De R_{1m}-toets kan echter geen onderscheid maken tussen die twee soorten schendingen. Een goede strategie is daarom, eerst de R_0-toets toe te passen en als er geen duidelijke schending is van de normaliteit gebruik te maken van de R_{1m}-toets. Men hoede zich echter voor een al te absolute interpretatie. Een significante R_{1m}-toets, samen met een niet significante R_0-toets is geen bewijs dat aan de assumptie van normaliteit is voldaan, en dat de modelschendingen dus bij het Raschmodel moeten liggen. Wil men deze twee assumpties duidelijk scheiden, dan verdient het de voorkeur de assumptie van normaliteit helemaal niet te maken, en CML als schattingsmethode te gebruiken.

4.3.6 Een voorbeeld

Als voorbeeld wordt een artificiële dataset geanalyseerd waarbij de itemantwoorden aan het Raschmodel voldoen, maar waarbij de verdeling van θ scheef is. De θ-waarden zijn gedefinieerd als

$$\theta = \frac{[\exp(-0.7\,z) - 1]}{-0.7}$$

waarbij z een aselecte trekking is uit de standaardnormale verdeling. De verdeling van θ is weergegeven in figuur 4.8, en wijkt dus sterk af van de normale verdeling. De toets bestaat uit 7 items met itemparameters (-1.5, -1, -0.5, 0, 0.5, 1, 1.5); n = 1000.



Figuur 4.8
Links scheve verdeling van θ

De schattingen en enkele statistische grootheden staan in tabel 4.11. De standaardfouten van de parameterschattingen zijn ongeveer 0.07. Vergeleken met deze grootte, verschillen CML- en MML-schattingen niet veel van elkaar.

Tabel 4.11

Schattingen en toetsingsgrootheden

item	$\hat{\beta}_i$ (CML)	$\hat{\beta}_i$ (MML)	S_i	vg	p	M_i	$M2_i$	$M3_i$
1	-1.460	-1.420	2.325	3	.508	0.76	1.50	0.84
2	-0.924	-0.933	0.817	3	.845	0.36	0.68	0.07
3	-0.506	-0.535	1.361	3	.715	-0.32	0.14	-0.15
4	0.053	0.021	2.853	3	.415	-0.94	-1.36	-0.97
5	0.394	0.371	1.255	3	.740	0.22	0.72	-0.04
6	0.964	0.972	7.288	3	.063	-2.60	-1.95	-1.65
7	1.480	1.526	6.752	3	.080	-2.52	0.61	2.30

R_{1c} = 19.17	vg = 18	p = .381
R_0 = 68.74	vg = 5	p < .00005
R_{1m} = 87.12	vg = 23	p < .00005

Voor de itemgerichte toetsen die in de tabel 4.11 zijn gerapporteerd is er niet veel reden om het model te verwerpen, hoewel voor de laatste twee items de overeenkomst met het model niet schitterend is. Vergelijken we dit echter met de uitkomsten van de *R*-toetsen, dan zien we dat de R_0- en de R_{1m}-toets zeer verschillende resultaten opleveren: de R_{1c}-toets, die niet beïnvloed wordt door de veronderstelling van de normale verdeling is niet significant. De conclusie is dus dat er geen reden is om het Raschmodel te verwerpen, maar een zeer overtuigende reden om deassumptie van een normale verdeling te verwerpen. In tabel 4.12 zijn de geobserveerde voorspelde scoreverdelingen weergegeven, waarbij het patroon van de afwijkingen niet erg duidelijk is. Het aantal geobserveerde nul-scores, bijvoorbeeld, is duidelijk groter dan verwacht, doch bij de daaropvolgende lage scores, 1 en 2, is de geobserveerde frequentie kleiner dan verwacht. Het patroon van afwijkingen tussen geobserveerde en voorspelde scorefrequenties hangt op een ingewikkelde manier af van de itemparameters en de verdeling van θ. In het algemeen is het niet mogelijk een duidelijke aanwijzing te krijgen over de onderliggende verdeling van θ door deze afwijkingen te bestuderen.

Tabel 4.12

Geobserveerde en verwachte scorefrequenties

score	geobs.	verwacht
0	98	61.3
1	94	131.1
2	147	180.2
3	188	197.4
4	212	180.9
5	176	137.6
6	72	81.3
7	13	29.7

Tenslotte zij er nog op gewezen dat, hoewel de assumptie van normaliteit op grove wijze geschonden is, de parameterschattingen met CML en MML erg goed op elkaar lijken. Het Raschmodel aangevuld met de normale verdeling voor θ is blijkbaar erg robuust tegen schendingen van de normaliteit. Men dient zich echter te hoeden voor klakkeloze generalisatie van dit resultaat. Een meer gedetailleerde studie is te vinden in Zwinderman (1991, hoofdstuk 4). In hoofdstuk 7 wordt een voorbeeld gegeven waarbij een verkeerde specificatie van de verdeling van θ leidt tot serieuze systematische fouten in de schatting van de itemparameters.

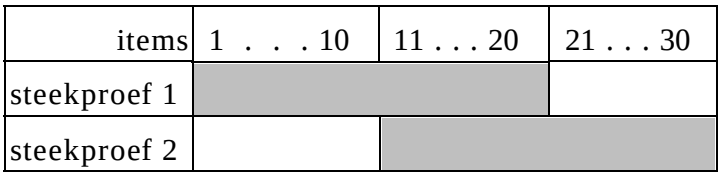
4.4 Het Raschmodel en onvolledige designs

In de vorige paragrafen is het Raschmodel uitvoerig besproken voor een situatie waarin alle personen uit de steekproef alle items beantwoorden. In de praktijk zal dit heel vaak niet het geval zijn, omdat sommigen door gebrek aan tijd de laatste items niet meer kunnen beantwoorden of omdat om een of andere reden bepaalde items worden overgeslagen. Het ontbreken van itemantwoorden in deze gevallen is dan afhankelijk van de persoon zelf die de items beantwoordt. De gaten die aldus in de data ontstaan zijn niet gepland. Analyse van zulke data is niet eenvoudig, en kan leiden tot systematische fouten in de parameter- schattingen, afhankelijk van de reden die tot het niet beantwoorden van bepaalde items heeft geleid. Als bijvoorbeeld items worden overgeslagen omdat ze moeilijk zijn, of er moeilijk uitzien, is het redelijk om aan te nemen dat de kans dat een item wordt overgeslagen groter is naarmate de vaardigheid waarop een beroep wordt gedaan lager is. In zo'n geval dient men uiterst voorzichtig te zijn met schattingsmethoden. Details hierover zijn het onderwerp van hoofdstuk 6.

Soms echter worden de gaten in de data gepland. Bij het construeren van een itembank van 1000 items zal het in de meeste gevallen om praktische redenen ondoenlijk zijn om alle personen alle items te laten beantwoorden. Daarom wordt aan elke persoon slechts een gedeelte van de items ter beantwoording voorgelegd volgens een vooropgezet design. In zo'n geval spreekt men van structureel onvolledige designs. De planning van een design kan echter verschillende vormen aannemen. Uitgaande van enige voorkennis over de moeilijkheidsgraad van de items zou een onderzoeker als volgt te werk kunnen gaan: aan de hand van een kleine voortoets van bijvoorbeeld 10 items die direct na afname nagekeken wordt, neemt men de beslissing voor de vervolgtoets. Personen met een lage score, zeg 5 of minder items juist, krijgen een relatief gemakkelijke natoets, de anderen een moeilijke natoets. Deze regel is eenduidig, maar er kan niet van te voren gezegd worden wie welke natoets zal krijgen. Het design staat dus onder de controle van degenen die de items beantwoorden. Daartegenover staat een design dat volledig van te voren is gepland. Bijvoorbeeld, de kinderen van school 1 krijgen toets 1, die van school 2 krijgen toets 2. Hier hebben de kinderen geen enkele controle op het design.

In deze paragraaf worden schattings- en toetsingsprocedures besproken die toepasbaar zijn in volledig door de onderzoeker gecontroleerde designs. De vraag welke procedures te gebruiken in andere gevallen, wordt in hoofdstuk 6 besproken.

In figuur 4.9 is een schematische weergave gegeven van een onvolledig design. De gearceerde oppervlakken stellen items voor die wel zijn aangeboden, de witte oppervlakken komen overeen met niet aangeboden items.



Figuur 4.9
Een onvolledig design met twee boekjes

Steekproef 1 heeft de items 1 tot 20 beantwoord en steekproef 2 de items 11 tot 30. Deze twee deelverzamelingen items worden doorgaans als een toetsboekje aangeboden, en om die reden zullen deelverzamelingen items die aan een groep personen worden aangeboden kortweg aangeduid worden als een boekje. Let wel dat in figuur 4.9 de boekjes elkaar overlappen.

In het algemeen zijn er B boekjes, en we definiëren de indexverzameling I_b $(b = 1, \cdots, B)$ als

$$I_b = \{i \mid \text{item } i \text{ komt voor in boekje } b\} \tag{4.103}$$

Het aantal items in boekje b wordt aangeduid als k_b. Het aantal personen dat boekje b heeft gekregen duiden we aan als n_b, en het aantal personen dat boekje b heeft gekregen en bovendien een score s (s = 0,...,k_b) heeft behaald, wordt aangeduid als n_{sb}. Een analoge notatie wordt ook gebruikt voor het aangeven van proporties en kansen. Zo betekent $p_{i|sb}$ de proportie juiste antwoorden op item i in de subgroep van personen die boekje b hebben gekregen en een score s hebben behaald.

Het totale aantal items dat in de analyse is betrokken duiden we aan met k. In figuur 4.9 geldt dus dat k=30. De antwoordvariabele X_i die bij volledige designs slechts twee waarden, 0 en 1, kon aannemen, laten we bij onvolledige designs drie waarden aannemen. We kennen X_i de waarde c toe indien het item niet is aangeboden, waarbij c een willekeurige waarde is die verschilt van 0 en 1. Voor een persoon met vaardigheid θ kunnen we nu twee conditionele kansverdelingen van X_i beschouwen, een voor het geval item i is aangeboden, en een voor het geval dat item i niet is aangeboden. Deze twee verdelingen zijn weergegeven in de rijen van tabel 4.12.

Tabel 4.12
Verdeling van X_i, conditioneel op θ en op het design

	$X_i = 0$	$X_i = 1$	$X_i = c$
aangeboden	$1 - f_i(\theta)$	$f_i(\theta)$	0
niet aangeboden	0	0	1

In de verdeling waarbij het item niet is aangeboden, kan X_i maar één waarde aannemen met een kans groter dan 0. In zo'n geval zegt men dat de verdeling van X_i gedegenereerd is. Formeel echter kunnen we de gewone algebra bedrijven met deze variabele en haar kans- verdeling.

Om expliciet aan te geven naar welke van de twee verdelingen we verwijzen voeren we de indicatorvariabelen D_{bi} in, die gedefinieerd zijn als

$$D_{bi} = \begin{cases} 1 \text{ indien } i \in I_b \\ 0 \text{ indien } i \notin I_b. \end{cases}$$

Eerst wordt de CML-schattingsprocedure besproken. Om het model te kunnen schrijven als een multinomiaal model moeten we de designvariabelen D_{bi} als toevalsvariabelen beschouwen. Dit kunnen we doen door voor de verschillende boekjes

een verzadigd multinomiaal model te beschouwen met parameters ω_b, de kans dat boekje b wordt aange-
boden. De ML-schatter van deze parameters is gegeven door

$$\hat{\omega}_b = \frac{n_b}{n}, \quad (b = 1, \cdots, B). \tag{4.104}$$

De multinomiale kans op een antwoordpatroon $\boldsymbol{x}$ is dan gegeven door

$$\begin{aligned} P(\boldsymbol{x}) &= P(\boldsymbol{x}|s,b)\, P(s,b) \\ &= P(\boldsymbol{x}|s,b)\, P(s|b)\, P(b) \\ &= \pi_{\boldsymbol{x}|sb}\, \omega_{s|b}\, \omega_b , \end{aligned} \tag{4.105}$$

waarbij de laatste regel niets anders is dan een verkorte notatie van de regel erboven. Voor de verdeling van de scores binnen een boekje nemen we, net als in het geval van een volledig design, een verzadigd multinomiaal model aan. De ML-schatters van de parameters van dit model zijn dus gegeven door

$$\hat{\omega}_{sb} = \frac{n_{sb}}{n_b}. \tag{4.106}$$

Gebruik makend van (4.104) en (4.106) zien we dus dat in (4.105) alleen de factor $\pi_{\boldsymbol{x}|sb}$ afhangt van de itemparameters, maar ook dat de conditie niet louter en alleen de score s is, maar de combinatie (s,b). Verzamelen we nu de itemparameters van alle items die behoren tot boekje b in de vector $\boldsymbol{\varepsilon}_b$, dan is $\pi_{\boldsymbol{x}|sb}$ gegeven door

$$\pi_{\boldsymbol{x}|sb} = \frac{\prod_{i=1}^{k} \varepsilon_i^{d_{bi}x_i}}{\gamma_s(\boldsymbol{\varepsilon}_b)} = \frac{\prod_{i\in I_b} \varepsilon_i^{x_i}}{\gamma_s(\boldsymbol{\varepsilon}_b)}. \tag{4.107}$$

De middelste uitdrukking in (4.107) geeft duidelijk aan hoe, door gebruik te maken van de waarde d_{bi} alle k antwoordvariabelen in de kansuitdrukking kunnen worden opgenomen, terwijl het rechterlid overeenkomt met het rechterlid van (4.40): het is gewoon de conditio- nele kans op het antwoordpatroon gegeven de score, maar beperkt tot de items die zijn aangeboden. Omdat in de totale steekproef alle antwoordpatronen onafhankelijk zijn van elkaar, is de aannemelijkheidsfunctie het produkt van

uitdrukkingen zoals het rechterlid van (4.107), en de log-aannemelijkheidsfunctie is de som van hun logaritmen.

Als dat duidelijk is, ligt de afleiding van de schattingsvergelijkingen, de uitdrukkingen voor de informatiematrix en de toetsingsgrootheden S_i^*, M_i^* en R_{1c}^* voor de hand. We geven ze hier volledigheidshalve, een gedetailleerde afleiding kan men vinden in Verhelst en Eggen (1989) en in Glas (1989).

De schattingsvergelijkingen zijn gegeven door

$$t_i = \sum_{b:i\in \mathrm{I}_b} \sum_{s=0}^{k_b} n_{sb} \frac{\varepsilon_i \gamma_{s-1}(\boldsymbol{\varepsilon}_b)}{\gamma_s(\boldsymbol{\varepsilon}_b)}, \tag{4.108}$$

waarin t_i het totaal aantal juiste antwoorden is dat op item i is uitgebracht.

De uitdrukkingen voor de informatiematrix zijn een veralgemening van (4.48):

$$I_{ij}(\beta) = \begin{cases} \sum_{b:i\in I_b} \sum_{s}^{k_b} n_{sb}[\pi_{i|s}(1-\pi_{i|s})] & \text{indien } i=j, \\ \sum_{b:i,j\in I_b} \sum_{s}^{k_b} n_{sb}[\pi_{ij|s} - \pi_{i|s}\pi_{j|s_v}] & \text{indien } i\neq j. \end{cases} \tag{4.109}$$

Voor de S_i-toetsen verandert er heel weinig. Het enige dat aangepast moet worden is de groepering van scores in scoregroepen G_q. Bij een volledig design konden we volstaan met het groeperen van scores; hier moeten de combinaties (s,b) gegroepeerd worden. De manier van groeperen is bepalend voor het onderscheidend vermogen van de toets tegen bepaalde schendingen van het model. Een concreet voorbeeld hiervan zal besproken worden in hoofdstuk 9 bij de discussie over itemonzuiverheid. De formule voor de benaderende grootheid S_i^* komt dan neer op een eenvoudige verandering van (4.98):

$$S_i^* = \sum_{q=1}^{r} \frac{\left[\sum_{(s,b)\in G_q} n_{sb}(p_{i|sb} - \hat{\pi}_{i|sb})\right]^2}{\sum_{(s,b)\in G_q} n_{sb}\hat{\pi}_{i|sb}(1-\hat{\pi}_{i|sb})}. \tag{4.110}$$

Voor de M-toetsen geldt precies hetzelfde: alle (s,b) combinaties worden opgedeeld in een laag- een midden- en een hoog-groep. Om die combinaties te ordenen moeten we echter beschikken over een ordeningsprincipe, dit wil zeggen we moeten een rationele methode vinden om alle combinaties (s,b) een rangnummer $w(s,b)$ te geven. In het programma OPLM worden de rangnummers zo toegekend dat

$$w(s,b) < w(s',b') \text{ indien } \hat{\pi}_{i|sb} < \hat{\pi}_{i|s'b'} \ . \tag{4.111}$$

Indien de twee geschatte kansen aan elkaar gelijk zijn beslist het toeval over de nummering. Op deze manier kunnen scores geordend worden, ook als ze afkomstig zijn van verschillende boekjes.

Bij de veralgemening van de R_{1c}-toets tot onvolledige designs treedt er een complicatie op. In paragraaf 4.3.5 werd gezegd dat de opdeling in scoregroepen voor alle items dezelfde moet zijn, omdat anders de Y matrix van de kwadratische vorm niet teruggebracht kan worden tot een blokdiagonale structuur. Bij onvolledige designs kan deze gelijke opdeling natuurlijk niet, want het ordeningsprincipe (4.111) is zinloos indien item i niet voorkomt in boekje b of b'. Daarom wordt een opdeling gemaakt per boekje in r_b scoregroepen G_{bq}, $(q=1,...,r_b)$, en de veralgemening van (4.101) is dan gegeven door

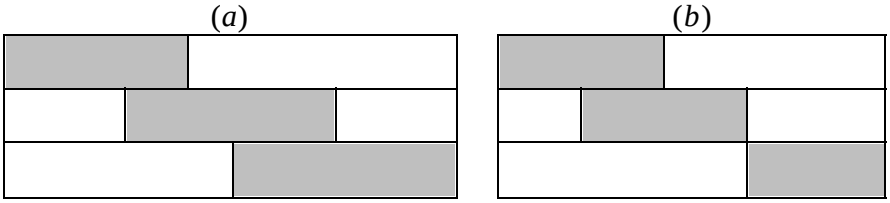
$$R_{1c}^{*} = \sum_{b} \sum_{q=1}^{r_b} \sum_{i \in I_b} \frac{\left[\sum_{s \in G_{bq}} n_s (p_{i|sb} - \hat{\pi}_{i|sb}) \right]^2}{\sum_{s \in G_{bq}} n_s \hat{\pi}_{i|sb} (1 - \hat{\pi}_{i|sb})} \ . \tag{4.112}$$

Het aantal vrijheidsgraden is gegeven door

$$\sum_{b=1}^{B} \left[r_b (k_b - 1) \right] - (k - 1) \ .$$

Hoewel de technische aspecten van het schatten van de parameters eigenlijk alleen neerkomen op iets meer gecompliceerde formules, waar een gebruiker bij zijn eigen toepassingen niet veel last van heeft, als programmatuur gebruikt wordt waar deze formules in zijn geïmplementeerd, is er een ander probleem waarmee bij het plannen van onderzoek terdege rekening moet worden gehouden. In figuur 4.9 zijn twee boekjes afgebeeld die overlappen. In zo'n geval zal men zeggen dat het design verbonden is. Bij ingewikkelder designs is de definitie van verbondenheid iets ingewikkelder. In figuur 4.10 zijn twee designs afgebeeld met elk drie boekjes. Het design (a) is verbonden,

hoewel boekje 1 en boekje 3 geen gemeenschappelijke items hebben, maar boekje 1 vertoont overlap met boekje 2, en boekje 2 heeft overlap met boekje 3, hoewel er geen enkel item is dat in alle drie de boekjes voorkomt. Het design (*b*) is niet verbonden want boekje 3 heeft geen enkele overlap met boekje 1 of boekje 2.



Figuur 4.10
Een verbonden (*a*) en een niet-verbonden design (*b*)

In een niet-verbonden design bestaan geen unieke CML-schatters van de itemparameters. Dit hoeft ook geen verwondering te wekken, omdat het nu eenmaal onmogelijk is om de relatieve moeilijkheid van twee items te schatten als niemand beide items heeft beantwoord. Willen we toch gegevens die verzameld zijn onder design (*b*) in figuur 4.10 met het Raschmodel analyseren, dan kan dat alleen door een MML-procedure te gebruiken.

Bij de MML-schattingsprocedure hebben we iets meer vrijheid om de verdeling van θ te specificeren dan bij volledige designs. In het design gegeven in figuur 4.9 bijvoorbeeld zou het kunnen zijn dat de twee steekproeven aselect uit dezelfde populatie zijn getrokken. In dat geval moeten naast de itemparameters de twee parameters van die gemeenschappelijke verdeling worden geschat. Het zou echter ook kunnen dat die twee steekproeven uit twee verschillende populaties zijn getrokken. Dan moeten, behalve de itemparameters, ook twee gemiddelden en twee varianties worden geschat. Voor het design (*a*) uit figuur 4.10 hebben we nog meer mogelijkheden: we kunnen een enkele verdeling veronderstellen, of twee of drie. Bij twee verdelingen zijn twee van de drie steekproeven afkomstig uit dezelfde populatie. In het algemeen kunnen we dus A populaties of verdelingen beschouwen, en uit elke populatie hebben we een of meer steekproeven die een boekje voorgelegd krijgen. Dus $A \leq B$, en er moeten $2A$ populatieparameters geschat worden: μ_a en σ_a^2, $(a = 1,\ldots,A)$. De log-aannemelijkheidsfunctie is dan een voor de hand liggende veralgemening van (4.58)

$$\ln L(\beta,\boldsymbol{\mu},\boldsymbol{\sigma^2};\boldsymbol{X}) = \sum_{b=1}^{B}\sum_{v=1}^{n_b} \ln \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta)\,\frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left[-\frac{(\theta-\mu_a)^2}{2\sigma_a^2}\right]\mathrm{d}\theta\,, \qquad (4.113)$$

waarin $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)$ en $\sigma^2 = (\sigma_1^2, \ldots, \sigma_A^2)$. De index *a* in (4.113) dient begrepen te worden als een functie van het boekjesnummer en dient dus gelezen te worden als $a(b)$, de populatie waaruit de steekproef, die boekje *b* heeft gekregen, afkomstig is.

Bij niet-verbonden designs is men niet helemaal vrij om steekproeven aan verschillende populaties toe te wijzen. In design (*b*) van figuur 4.10, bijvoorbeeld, kan men wel een analyse uitvoeren met de hypothese van één of twee verschillende populaties, maar in de tweede geval kan men niet veronderstellen dat steekproef 1 en 2 afkomstig zijn uit dezelfde populatie en steekproef 3 uit een andere. Veronderstelt men echter dat steekproef 1 en steekproef 3 uit dezelfde populatie komen, dan zijn alle parameters in principe wel schatbaar, omdat de items uit die twee boekjes met elkaar verbonden worden door een gemeenschappelijke verdeling.

Tot slot van deze paragraaf, nog een opmerking over schatbaarheid van parameters in het algemeen. Als gezegd wordt dat voor het design in figuur 4.9 CML-schatters bestaan, dan betekent dit niet dat in alle gevallen waar dit design wordt toegepast CML-schattingen kunnen worden gevonden. Het zou bijvoorbeeld kunnen voorkomen dat in een bepaalde steekproef een item door iedereen juist beantwoord is. In zo'n geval bestaat er geen eindige CML-schatting voor de parameter van dit item. Bij onvolledige designs zijn de voorwaarden waar- onder eindige en unieke CML-schattingen van de parameters bestaan echter veel ingewikkelder dan het voorbeeldje hiervoor suggereert. Algemene voorwaarden, die ook redelijk gemakkelijk met de computer kunnen gecontroleerd worden, zijn gegeven in Fischer (1981) en worden in hoofdstuk 6 besproken. Voor het bestaan van MML-schattingen zijn de algemene voorwaarden niet precies bekend. In het algemeen zijn die voorwaarden echter milder dan voor CML-schattingen: als CML-schattingen bestaan, bestaan ook MML-schattingen; maar MML-schattingen kunnen ook bestaan waar CML onmogelijk is. Design (*b*) uit figuur 4.10 is daar een voorbeeld van.

4.5 Het schatten van de persoonsparameters

Het uiteindelijke doel bij het ontwikkelen van een meetinstrument is het meten van eigenschappen van objecten of personen, dat wil zeggen het toekennen van getallen aan die objecten of personen zodanig dat de toegekende getallen ook de mate van aanwezigheid van de bedoelde eigenschap aangeven. In de context van het Raschmodel betekent dit de waarde van θ 'berekenen' voor een willekeurige persoon. De observaties die we nodig hebben, zijn de itemantwoorden van die persoon. De waarde van θ is dus een functie van de itemantwoorden. Als we een toets tweemaal afnemen

aan dezelfde persoon, zullen de item-antwoorden niet tweemaal dezelfde zijn. Itemantwoorden zijn dus toevalsvariabelen, en bijgevolg is de waarde van θ die we uit deze antwoorden berekenen ook een toevalsvariabele. Vergelijk met lichaamslengte: de observatie die we nodig hebben om lichaamslengte te bepalen is iemands verticale uitgestrektheid en die varieert ook: na een dag vol activiteiten is iemands verticale uitgestrektheid minder dan na een nacht slaap. Het is dus niet zonder meer duidelijk wat bedoeld wordt met lichaamslengte: ook als we de observatie-omstandigheden standaardiseren (bijvoorbeeld altijd 's morgens na minstens zes uur rust), zullen de meetuitslagen variabiliteit vertonen, en als we slechts een keer meten, weten we niet of we een 'lage' dan wel een 'hoge' uitkomst hebben. Meestal maken we ons echter niet druk over dit probleem omdat voor de praktische bedoelingen waar we deze uitkomsten voor nodig hebben, de variabiliteit van de uitkomsten te verwaarlozen is. Bij het meten van schoolse of cognitieve vaardigheden met de meetinstrumenten waarover we beschikken, is die variabiliteit meestal niet te verwaarlozen. We zullen er dus enige aandacht aan moeten besteden.

Er zijn bovendien nog twee overwegingen van technische aard waar men rekening mee moet houden bij de interpretatie van de berekende θ-waarde, namelijk de normalisering van de itemparameters en de toegepaste rekenregel. We illustreren beide wederom aan de hand van het voorbeeld over lichaamslengte.

Gewoonlijk bedoelen we met lichaamslengte de afstand tussen iemands voetzolen en kruin bij gestrekte houding. De eenheid waarin we meten wordt gewoonlijk toegevoegd aan de meetuitslag. Zo spreken we van een lichaamslengte van 176 cm of 69 inch. Bij het meten van vaardigheden worden meestal geen eenheden toegevoegd, doch zoals uiteengezet in paragraaf 4.3.1 is er wel degelijk van een eenheid sprake die we kunnen kiezen: de waarde van de gemeenschappelijke discriminatieparameter is willekeurig en bepaalt de eenheid waarin we meten. Als twee meetuitslagen met elkaar worden vergeleken, dienen we er dus zeker van te zijn dat ze in dezelfde eenheid zijn uitgedrukt. Een analoog argument geldt ook voor het nulpunt van de schaal. We zouden iemands lichaamslengte ook kunnen definiëren als de afwijking tot het populatiegemiddelde of het aantal centimeters dat hij in rechtopstaande houding uitsteekt boven een tafel van één meter hoog. Het nulpunt van de schaal wordt bepaald door wat we de normalisatie genoemd hebben. Twee meetuitslagen zijn dus alleen zinvol te vergelijken als ze afkomstig zijn van twee meetinstrumenten met hetzelfde nulpunt en dezelfde eenheid.

Het belang van de rekenregel kan als volgt geïllustreerd worden voor het voorbeeld van de lichaamslengte. Voor het bepalen van iemands lichaamslengte laten we tien beoordelaars een 'schatting-op-zicht' van de lichaamslengte maken. Als eerste

rekenregel nemen we het gemiddelde van de tien schattingen. Bij de tweede rekenregel verwijderen we eerst de hoogste en de laagste schatting en we nemen als uitkomst het gemiddelde van de acht overblijvende schattingen. Het is duidelijk dat we bij het bepalen van iemands lichaamslengte volgens de twee rekenregels, in het algemeen twee verschillende uitkomsten zullen krijgen. Bovendien is het niet meteen duidelijk welke de beste regel is: de eerste regel is iets nauwkeuriger dan de tweede omdat hij gebaseerd is op tien schattingen en de tweede slechts op acht. Daartegenover staat echter dat de tweede regel beter beschermd is tegen grove vergissingen van de beoordelaars. Voor de schattingen van de vaardigheden hebben we ook verschillende rekenregels, die verschillende uitkomsten geven. Welke rekenregel we moeten kiezen is afhankelijk van het gebruik van de meetresultaten. Omdat hieraan soms serieuze ethische implicaties verbonden zijn, zullen we tamelijk uitvoerig op deze regels ingaan.

In paragraaf 4.5.1 worden de verschillende rekenregels besproken. Omdat elke regel een schatting van θ geeft worden die regels gewoonlijk aangeduid als schattingsmethode. Paragraaf 4.5.2 behandelt een voorbeeld.

Bij de bespreking van de veronderstellingen die aan het Raschmodel ten grondslag liggen, is er op gewezen dat homogeniteit met betrekking tot het Raschmodel wordt verondersteld. Dit betekent dat er van uit gegaan wordt dat het Raschmodel voor iedere persoon in de steekproef geldt, of, indien er schendingen zijn van de axioma's, dat die schendingen in gelijke mate voor iedere persoon gelden. Nu is het natuurlijk mogelijk dat het Raschmodel geldt voor de overgrote meerderheid van de personen in de steekproef, maar voor een enkeling niet. In zo'n geval is het goed mogelijk dat dit gebrek aan homogeniteit niet ontdekt wordt door de statistische toetsen die in paragraaf 4.3 werden besproken. Door individuele antwoordpatronen nader te onderzoeken kan men soms overtuigende evidentie vinden dat in individuele gevallen het Raschmodel als nulhypothese verworpen moet worden. Dit is het onderwerp van paragraaf 4.5.3.

4.5.1 Drie methoden om de persoonsparameter θ te schatten

De drie methoden die we hier bespreken, worden aangeduid als ML, Warm of WML en EAP, en staan respectievelijk voor 'Maximum Likelihood', 'Weighted Maximum Likelihood' en 'Expected A Posteriori'. The WML-methode is ontwikkeld door Warm (1989). Vooraleer we de verschillende methoden uiteenzetten, is het belangrijk te wijzen op een overeenkomst in de drie methoden. Om θ te schatten, moeten we de waarde van de itemparameters kennen. In de praktijk kennen we die natuurlijk nooit, en daarom

gebruiken we geschatte waarden. Bij het schatten van θ wordt gedaan alsof die geschatte waarden van de itemparameters de echte waarden zijn. Daarmee wordt dus een extra fout geïntroduceerd in de schatting van θ. Hoe erg die fout is, hangt af van de standaardfout van de itemparameterschattingen, en deze hangt op haar beurt weer in belangrijke mate af van de grootte van de calibratiesteekproef. In het gebruik wordt echter zelden met die fout rekening gehouden, er wordt gedaan alsof die fout er niet is, waardoor de nauwkeurigheid van de θ-schatting doorgaans overschat wordt. Het precieze onderzoek naar de invloed van die schattingsfout op de nauwkeurigheid van de schatting van θ is nogal moeilijk, en wordt hier verder niet besproken.

De ML-schatter van θ

Indien de itemparameters bekend zijn, en we observeren één antwoordpatroon $\boldsymbol{x}$, dan is de logaritme van de aannemelijkheidsfunctie gegeven als een speciaal geval van (4.28):

$$\ln L(\theta;\boldsymbol{x},\beta) = s\theta + \sum_{i=1}^{k} x_i(-\beta_i) - \sum_{i=1}^{k} \ln[1+\exp(\theta-\beta_i)], \qquad (4.114)$$

waarin $s = \Sigma_i x_i$ de score is. Merk op dat in (4.114) de itemparameters β_i als constanten worden behandeld: de tweede term in het rechterlid is dus uitsluitend een functie van de data. De derde term is alleen functie van de parameter θ, zodat duidelijk is dat (4.114) de gedaante heeft van een log-aannemelijkheidsfunctie in de exponentiële familie. De schattings- vergelijking is dus onmiddellijk gegeven door

$$s = \mathscr{E}(S) = \sum_{i=1}^{k} \mathscr{E}(X_i) = \sum_{i=1}^{k} f_i(\theta)\ . \qquad (4.115)$$

Hoewel de formule erg eenvoudig is, is voor het berekenen van de waarde van θ een iteratieve procedure nodig; een expliciete oplossing bestaat niet. De meeste computer-programmatuur geeft de oplossingen echter standaard. Vergelijking (4.115) heeft echter niet altijd een oplossing. Omdat $0 < f_i(\theta) < 1$ is het rechterlid van (4.115) altijd groter is dan 0 en altijd kleiner dan de maximale toetsscore k. Voor de scores 0 en k is er dus geen enkele waarde van θ waarvoor aan (4.115) voldaan is. Voor alle andere scores bestaat de ML-schatting wel. Men dient dus voorzichtig te zijn bij het berekenen van

steekproefgrootheden, zoals de gemiddelde ML-schatting. Het invullen van een willekeurige lage θ-waarde voor personen met een nul-score en een willekeurige hoge waarde in geval van perfecte scores is uit den boze. Wil men toch per se een gemiddelde berekenen, dan zit er niets anders op dan personen met zulke extreme scores uit de steekproef te verwijderen, maar daardoor kunnen groepsvergelijkingen onzuiver gaan worden. Stel dat in een steekproef 5% perfecte scores voorkomen. Hoewel er geen ML-schattingen bestaan voor die 5%, weten we toch dat we de vaardigheid van die personen hoog moeten inschatten. Door ze te verwijderen gaan we de gemiddelde vaardigheid in die steekproef, en bij veralgemening dus ook in de geassocieerde populatie, onderschatten. Komen in een andere steekproef (uit een andere populatie) slechts 2% perfecte scores voor, dan treedt er ook een onderschatting op, maar die is minder erg. De twee berekende gemiddelden kunnen dan niet zinvol met elkaar worden vergeleken.

De nauwkeurigheid waarmee θ gemeten wordt is de nauwkeurigheid waarmee θ geschat wordt en deze kan, zoals in paragraaf 4.2.1 werd uiteengezet, worden afgeleid uit de informatiefunctie, die hier de naam toetsinformatiefunctie draagt:

$$I(\theta) = \sum_{i=1}^{k} f_i(\theta)[1-f_i(\theta)]. \qquad (4.116)$$

Het produkt $f_i(\theta)[1-f_i(\theta)]$ bereikt zijn grootste waarde indien $f_i(\theta) = 0.5$, en dit is het geval indien $\theta = \beta_i$. Dit produkt wordt kleiner naarmate θ verder afwijkt van β_i. Vullen we nu in (4.116) een waarde in die ver afligt van alle β's, dan blijkt dat de toets zeer weinig informatie oplevert over die θ. Indien de waarde van θ middenin tussen de β's is gelegen, levert de toets meer informatie op over θ. Een toets kan dus voor bepaalde personen zeer informatief zijn, en voor andere niet. Deze geschiktheid wordt ook weerspiegeld in de standaardfout van de schatting van θ:

$$SE(\hat{\theta}) \approx \sqrt{1/I(\theta)}\ . \qquad (4.117)$$

Om (4.117) te evalueren moet men θ kennen. In een concrete toepassing waarbij men θ gewoonlijk niet kent, vult men in het rechterlid de ML-schatting van θ in. Het resultaat is natuurlijk een schatting van de standaardfout. Bovendien zijn rechter- en linkerlid van (4.117) slechts asymptotisch aan elkaar gelijk, dus indien $k \to \infty$. In toepassingen met een klein aantal items moet er rekening mee worden gehouden dat gebruik van (4.117) een forse onderschatting van de standaardfout kan opleveren.

De ML-schatter van θ heeft nog een tweede nadeel naast het feit dat hij niet bestaat voor perfecte en nulscores. Hij is namelijk zeer onzuiver. Het begrip zuiverheid dient

als volgt opgevat te worden. Stel dat een persoon met een bepaalde waarde θ een gegeven toets een zeer groot aantal keren maakt, in de veronderstelling van volledige 'brain wash' na elke afname, dan verwachten we niet dat hij telkens dezelfde score haalt. We zullen dus een verdeling van scores vinden. Als we even de gevallen waarin hij 0 of een perfecte score haalt buiten beschouwing laten, kunnen we voor elke score de ML-schatting berekenen. We beschikken dus ook over de verdeling van ML-schattingen. Een schatter heet zuiver als het gemiddelde van die verdeling gelijk is aan de echte θ-waarde. De afwijking tussen het gemiddelde van die verdeling en de echte waarde wordt de onzuiverheid of bias genoemd: bias = $\mathscr{E}(\hat{\theta}|\theta) - \theta$. De ML-schattingen zijn onzuiver in een heel speciale zin. Voor kleine waarden van θ is de onzuiverheid negatief en voor grote waarden positief. Wat precies bedoeld wordt met groot en klein is nogal ingewikkeld, doch in grote lijnen komt het op het volgende neer: meestal is de toetsinformatiefunctie ééntoppig, dat wil zeggen dat de infor- matie heel klein is voor zeer kleine waarden van θ, toeneemt tot een bepaalde θ-waarde, zeg θ_0, en vanaf daar weer afneemt. Met klein wordt nu grofweg bedoeld kleiner dan θ_0, en met groot, groter dan θ_0. Bovendien neemt de onzuiverheid toe naarmate θ meer van θ_0 afwijkt. Het effect van die onzuiverheid is dus als het ware een uitrekken van de schaal van geschatte θ's in vergelijking met de schaal van de echte θ's (zie Lord, 1983a, voor een gedetailleerde uiteenzetting).

Samenvattend: de ML-schatter van θ bestaat niet voor perfecte en nulscores, en is behoorlijk onzuiver. Dit zijn voldoende redenen om die schatter niet te gebruiken. Hij is in de literatuur vrij lang gebruikt omdat er geen goed alternatief was. Warm heeft in 1989 een θ-schatter ontwikkeld die beide euvels verhelpt. Die schatter wordt in de volgende paragraaf besproken.

De WML-schatter van θ (Warm-schatter)

Warm (1989) heeft aangetoond dat de onzuiverheid van de θ-schatter grotendeels kan worden opgeheven door niet de aannemelijkheidsfunctie te maximaliseren, maar een gewogen aannemelijkheidsfunctie. (WML staat voor Weighted Maximum Likelihood.) In het Raschmodel is deze weegfunctie de vierkantswortel uit de informatiefunctie. De WML-schatting van θ is dus die waarde van θ die de functie

$$W(\theta) = L(\theta)\sqrt{I(\theta)} \qquad (4.118)$$

maximaliseert.

De WML-schatter vertoont bijna geen onzuiverheid meer, tenzij voor zeer extreme θ-waarden. De overblijvende onzuiverheid vertoont daarenboven het omgekeerde beeld van de onzuiverheid voor de ML-schatter. Voor zeer kleine waarden van θ is de onzuiverheid positief, en voor zeer grote waarden negatief. De schaal van de geschatte θ's (met WML) vertoont dus een zekere krimping in vergelijking met de echte θ-waarden.

Een gelukkige bijkomstigheid van de WML-schatter is dat hij altijd bestaat, ook voor perfecte en nulscores.

De WML-schatter, samen met een schatting van de standaardfout en een schatting van de bias, wordt berekend in het programmapakket OPLM. De formule voor de standaardfout is ingewikkelder dan in het geval van de ML-schatter en wordt hier niet besproken.

De EAP-schatter van θ

Bij de ML- en de WML-schatter wordt alleen gebruik gemaakt van het geobserveerde antwoordpatroon om θ te schatten. Twee personen met dezelfde score behalen steeds dezelfde schatting van θ. Men zou echter ook andere informatie kunnen gebruiken om θ te schatten, bijvoorbeeld kennis omtrent de populatie waaruit de betrokken persoon afkomstig is. Dit is wat er gebeurt bij de EAP-schatter: daarin wordt informatie die men heeft over de populatie waaruit de betrokken persoon afkomstig is, gecombineerd met informatie die het antwoordpatroon oplevert. Deze combinatie levert in de regel een uitkomst op die ligt tussen de ML-schatting en het populatiegemiddelde. Bijvoorbeeld, stel dat men weet dat een persoon aselect uit een θ-populatie is getrokken en dat de gemiddelde θ-waarde in die populatie 0 is en de standaarddeviatie 1. Stel dat die persoon een hoge toetsscore haalt, met een ML-schatting van 3. Op grond van de toetsuitslag alleen zouden we besluiten tot een vaardigheids- schatting van 3, doch het veel lager gemiddelde van de populatie suggereert dat dit overdreven is. Immers, de kans dat er aselect een persoon met een θ-waarde van 3 of hoger wordt getrokken is zo klein, dat zich als het ware een correctie op de ML-schatter in de richting van het populatiegemiddelde opdringt. De EAP-schatter kan dus beschouwd worden als een soort compromis tussen de informatie die de toetsafname oplevert en de informatie over de populatie waarover we beschikken, net zoals de formule van Kelley die in hoofdstuk 3 is besproken.

Formeel is de EAP-schatter het gemiddelde van de a posteriori verdeling van θ, dit wil zeggen, de verdeling van θ indien de observaties gecombineerd worden met de a

priori verdeling van θ. Deze laatste verdeling is niets anders dan de verdeling van θ die aan het Raschmodel is toegevoegd om MML-schattingen te kunnen maken. De formules voor deze schatter volgen rechtstreeks uit het theorema van Bayes:

$$h(\theta|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\theta)g(\theta)}{P(\boldsymbol{x})}$$

$$= \frac{P(\boldsymbol{x}|\theta)g(\theta)}{\int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)g(\theta)\,d\theta}, \qquad (4.119)$$

waarbij de tweede gelijkheid rechtstreeks uit (4.56) volgt. De functie $h(\theta|\boldsymbol{x})$ is de a posteriori dichtheid van θ, waarbij duidelijk te zien is dat deze functie afhankelijk is zowel van de data en de itemparameters, via $P(\boldsymbol{x}|\theta)$, als van de a priori verdeling en de daarmee geassocieerde parameters, via $g(\theta)$. Het gemiddelde van de a posteriori verdeling is dan gegeven door

$$\mathscr{E}(\theta|\boldsymbol{x}) = \int_{-\infty}^{+\infty} \theta\, h(\theta|\boldsymbol{x})\,d\theta. \qquad (4.120)$$

De schatter zegt dus eigenlijk dat de persoon beschouwd dient te worden als een aselecte trekking uit een populatie van θ-waarden met dichtheidsfunctie $h(\theta|\boldsymbol{x})$. De schatter zelf is het gemiddelde van die populatie. Daaruit volgt geenszins dat twee personen met hetzelfde antwoordpatroon ook dezelfde θ-waarde hebben. Immers de a posteriori verdeling heeft ook een variantie ongelijk 0. Deze variantie, of de vierkantswortel eruit, de a posteriori standaarddeviatie, kan dus gehanteerd worden als een maat van onzekerheid. Deze variantie is gegeven door

$$\mathrm{var}(\theta|\boldsymbol{x}) = \int_{-\infty}^{+\infty} \theta^2 h(\theta|\boldsymbol{x})\,d\theta - [\mathscr{E}(\theta|\boldsymbol{x})]^2. \qquad (4.121)$$

De term 'expected a posteriori' is afkomstig uit de bayesiaanse statistiek. 'Echte' Bayesianen voeren de a priori verdeling, zowel de vorm, bijvoorbeeld de normale verdeling, als de waarde van de parameters, op als een soort geformaliseerde overtuiging. Bij toepassingen met MML-schattingen wordt alleen de vorm van de verdeling ingevoerd als hypothese, terwijl de parameters uit de data worden geschat. Deze benadering wordt aangeduid als empirisch bayesiaans. Bij de EAP-schattingsprocedure worden dus de geschatte populatieparameters gebruikt om de a priori verdeling te specificeren.

Stel nu dat men bij de schatting van de item- en populatieparameters twee steekproeven, afkomstig uit twee verschillende populaties, heeft gebruikt, die dezelfde toets hebben gekregen. Eénzelfde antwoordpatroon zal leiden tot verschillende EAP-schatters voor beide populaties, en wel in die zin dat de EAP-schatter voor een persoon uit de populatie met het laagste gemiddelde kleiner zal zijn dan voor een persoon uit de andere populatie. Indien men schattingen van θ gebruikt om beslissingen te nemen die individuen raken, dient men zich terdege bewust te zijn van de ethische implicaties bij het gebruik van EAP-schatters. Immers, de beslissing wordt niet uitsluitend gebaseerd op de itemantwoorden, doch ook op achtergrondinformatie, waarvan het gebruik in bepaalde contexten discriminerend of onrechtvaardig kan zijn. De beslissing om ze dan maar niet te gebruiken is echter een beetje simplistisch. Als men ze niet gebruikt is men aangewezen op ML- of WML-schatters, waarvan de standaardfout in de regel groter is dan de a posteriori standaarddeviatie, en grotere standaardfouten betekenen automatisch meer verkeerde beslissingen. Een goed gefundeerde verhandeling over dit onderwerp ontbreekt echter nog in de psychometrische literatuur.

4.5.2 Een voorbeeld

Als illustratie bij het commentaar dat in de vorige paragraaf gegeven is, beschouwen we het volgende artificiële voorbeeld. Veronderstel dat er twee populaties, A en B zijn waarin de vaardigheid normaal verdeeld is met een standaarddeviatie gelijk aan 1. Het gemiddelde van populatie A is -0.6 en dat van populatie *B* is +0.6. Uit beide populaties wordt aselect een steekproef getrokken van 250 personen. De toets die aan beide steekproeven wordt voorgelegd bestaat uit 21 Raschitems met parameters -2.0, -1.8, -1.6,...,1.6, 1.8, 2.0. De parameters worden geschat met CML, en vervolgens wordt voor ieder antwoordpatroon de ML- en de WML-schatter berekend. Daarnaast zijn ook MML-schatters berekend, waarbij naast de itemparameters ook twee gemiddelden en twee varianties worden geschat. Na de parameterschattingen zijn de schattingen van θ berekend volgens de drie methodes: ML, WML en EAP. Voor WML en EAP geldt, net als voor ML-schatters, dat de schatting alleen afhankelijk is van de score. De resultaten staan in tabel 4.13.

De getallen tussen haakjes in tabel 4.13 zijn de a posteriori standaarddeviaties (voor MML) of de standaardfouten (voor WML en ML). Omdat populatie B gemiddeld vaardiger is krijgen leden uit populatie B ook systematisch een hogere θ-schatting dan leden van populatie A voor dezelfde score. De a posteriori standaarddeviaties zijn ook systematisch kleiner dan de standaardfouten van de WML- en de ML-schatters. De

toets bereikt haar maximale informatie voor θ in de buurt van 0, en we zien ook dat de standaardfouten van WML en ML hun kleinste waarde bereiken rond dit punt. De plaats waar de a posteriori standaarddeviatie haar kleinste waarde bereikt is niet alleen afhankelijk van de informatiefunctie maar ook van de waarde van het gemiddelde en de standaarddeviatie, dus van de a priori verdeling. Merk tenslotte nog op dat de ML-schattingen meer 'uitgerekt' zijn dan de WML-schattingen, terwijl de EAP-schattingen meer samengedrukt zijn.

Tabel 4.13
EAP-, WML- en ML-schattingen van θ

score	EAP (pop. A)		EAP (pop. B)		WML		ML	
0	-3.194	(.574)	-2.748	(.532)	-4.416	(1.844)	---	---
1	-2.883	(.544)	-2.477	(.510)	-3.210	(.966)	-3.570	(1.052)
2	-2.600	(.520)	-2.227	(.492)	-2.590	(.757)	-2.769	(.781)
3	-2.341	(.500)	-1.991	(.479)	-2.141	(.658)	-2.251	(.669)
4	-2.098	(.485)	-1.768	(.467)	-1.773	(.601)	-1.848	(.606)
5	-1.870	(.473)	-1.553	(.459)	-1.453	(.565)	-1.505	(.568)
6	-1.651	(.463)	-1.346	(.453)	-1.161	(.541)	-1.198	(.542)
7	-1.441	(.455)	-1.143	(.448)	-.888	(.524)	-.914	(.525)
8	-1.236	(.450)	-.944	(.444)	-.628	(.513)	-.645	(.514)
9	-1.035	(.446)	-.748	(.443)	-.367	(.507)	-.385	(.507)
10	-.838	(.443)	-.551	(.443)	-.127	(.504)	-.130	(.504)
11	-.641	(.443)	-.355	(.443)	.120	(.504)	.123	(.504)
12	-.444	(.444)	-.157	(.446)	.369	(.507)	.379	(.507)
13	-.247	(.445)	.044	(.451)	.622	(.514)	.640	(.515)
14	-.048	(.449)	.249	(.456)	.884	(.526)	.910	(.526)
15	.156	(.454)	.460	(.463)	1.159	(.542)	1.196	(.544)
16	.364	(.460)	.679	(.473)	1.453	(.567)	1.505	(.569)
17	.579	(.469)	.909	(.485)	1.776	(.603)	1.850	(.608)
18	.804	(.480)	1.152	(.501)	2.146	(.660)	2.255	(.670)

19	1.041	(.494)	1.412	(.520)	2.597	(.758)	2.775	(.782)
20	1.293	(.511)	1.694	(.544)	3.219	(.967)	3.578	(1.053)
21	1.565	(.533)	2.006	(.574)	4.425	(1.845)	---	---

4.5.3 Passingsindices voor individuele antwoordpatronen

In de vorige paragraaf is gesteld dat de schatter van θ alleen afhankelijk is van de score. Dit kan enigszins paradoxaal klinken. Stel dat van twee personen die precies de helft van de items juist hebben beantwoord, de eerste de *k*/2 gemakkelijkste items juist had, en de tweede de k/2 moeilijkste. Is het dan niet redelijk de vaardigheid van de tweede hoger te schatten ? De oplossing van deze paradox is gelegen in het dubbele standpunt dat men tegenover statistische gegevens kan innemen. Statistische gegevens veronderstellen bij analyse steeds een model. Een gedeelte van de informatie die de gegevens bevatten gebruikt men voor het schattingsprobleem. Men kan de schattingen gebruiken en interpreteren, en de juistheid van de interpretatie is alleen gegarandeerd als de modelveronderstellingen juist zijn. Of deze veronderstellingen juist zijn weet men nooit met absolute zekerheid, doch men kan de juistheid statistisch toetsen door het gebruik van andere informatie in de data. In het gegeven voorbeeld is het inderdaad terecht aan beide personen dezelfde schatting van θ toe te kennen indien het model juist is. Beide antwoordpatronen zijn echter in een bepaalde betekenis vrij extreem, zodat men er aan kan twijfelen of de antwoorden wel volgens het Raschmodel tot stand zijn gekomen. Naast de vaardigheid kunnen natuurlijk tal van andere factoren het gedrag bepaald hebben, en de invloed van deze factoren kan zo belangrijk zijn dat het Raschmodel niet meer geldig is.

Redenen voor niet-passing van het model voor individuele respondenten kunnen bijvoor- beeld zijn: vermoeidheid, oneerlijk gedrag, systematisch verkeerd invullen van schrapkaarten waarbij het antwoord voor item *i* wordt ingevuld op de plaats *i*+1, enzovoort. Een discussie van deze en nog andere redenen voor systematische afwijkingen van het model kan men vinden in Hulin, Drasgow en Parsons (1983), die ook een groot aantal indices bespreken waarmee niet-passende antwoordpatronen ontdekt kunnen worden. Een recente en heel interessante bijdrage op dit gebied kan men ook vinden in Klauer (1991). Bij wijze van voorbeeld bespreken we hier een zeer eenvoudige index, die we aanduiden als $z(\theta,\boldsymbol{x})$:

$$z(\theta,\boldsymbol{x}) = \sum_{i=1}^{k} [\, f_i(\theta) - x_i \,] \; . \qquad (4.122)$$

De interpretatie van (4.122) is eenvoudig: hij geeft de som van de afwijkingen tussen het itemantwoord x_i en de verwachte waarde $f_i(\theta)$, elke term is dus verschillend van 0. Grote absolute afwijkingen ontstaan indien een verkeerd antwoord wordt gegeven bij gemakkelijke items of een juist antwoord bij moeilijke items. Indien een volmaakt Guttmanpatroon optreedt waarbij de *s* gemakkelijkste items juist worden beantwoord en de *k*-*s* moeilijkste fout, zijn de absolute afwijkingen relatief klein. Bij een antwoordpatroon waarbij het omgekeerde het geval is, krijgen we wel grote absolute afwijkingen, doch hun teken is verschillend: juiste antwoorden op een moeilijk item resulteren in een negatieve afwijking en verkeerde antwoorden op een gemakkelijk item geven een positieve afwijking, met als gevolg dat die in de som tegen elkaar zullen wegvallen, en kunnen resulteren in een kleine waarde van de index, net zoals bij een Guttmanpatroon. Deze index is dus niet erg geschikt. Een index die wel onderscheid maakt tussen Guttmanpatronen en hun omgekeerde is de ζ_2-index van Tatsuoka (1984):

$$\zeta_2(\theta,\boldsymbol{x}) = \frac{\sum_{i=1}^{k}[f_i(\theta)-x_i][f_i(\theta)-\bar{f}(\theta)]}{\left(\sum_{i=1}^{k} f_i(\theta)[1-f_i(\theta)][f_i(\theta)-\bar{f}(\theta)]^2\right)^{\frac{1}{2}}} \qquad (4.123)$$

waarin $\bar{f}(\theta) = \Sigma_i f_i(\theta)/k$. De interpretatie van ζ_2 is het gemakkelijkst indien we veronderstellen dat de items geordend zijn volgens oplopende moeilijkheid, en de score s ongeveer *k*/2 bedraagt. Voor een Guttmanpatroon waarbij de *s* makkelijkste items juist zijn beantwoord, zullen de eerste *s* termen van (4.123) overwegend negatief zijn, want $f_i(\theta)$-$x_i < 0$ voor $i<s$, en als de verdeling van de moeilijkheidsparameters niet al te scheef is, zal gelden dat $f_i(\theta) > \bar{f}(\theta)$ voor het merendeel van de eerste *s* items. Een omgekeerd Guttman- patroon zal resulteren in een positieve index. Bovendien kan aangetoond worden dat de verwachte waarde van ζ_2 gelijk is aan 0 en de variantie gelijk aan 1. Indien *k* niet al te klein is kan ζ_2 geïnterpreteerd worden als een standaardnormaal verdeelde variabele: waarden van de index groter dan 2 in absolute waarde hebben een kleine kans om geobserveerd te worden indien de nulhypothese, het Raschmodel, waar is. De ζ_2-index is in het programma OPLM geïmplementeerd.

Bij het interpreteren van deze indices dient men de nodige voorzichtigheid aan de dag te leggen. Indien de index gebruikt wordt om beslissingen te nemen die

verstrekkende gevolgen kunnen hebben voor een bepaalde persoon, dient men te bedenken dat het voorkomen van een ongebruikelijk of vreemd antwoordpatroon geen waterdicht bewijs is van bijvoorbeeld oneerlijk gedrag. Immers, indien men toetst met een significantieniveau van 5%, dan kan men verwachten dat ongeveer 5% van de antwoordpatronen in de steekproef een significante index zal opleveren indien het model juist is. Is dit percentage in de steekproef substantieel groter, dan wijst dit er op dat er iets niet in de haak is met het model. Nader onderzoek kan dan gewenst zijn, doch de index op zichzelf is een zwakke basis om individuele beslissingen te rechtvaardigen. Hij kan hoogstens leiden tot een grotere voorzichtigheid. De Amerikaanse naam voor dit soort indices, caution indices, is dan ook heel terecht.

Op deze en vele andere indices die in de literatuur zijn gebruikt hebben Molenaar en Hoijtink (1990) vanuit statistisch standpunt nogal wat kritiek geleverd. Deze kritiek komt erop neer dat we, om deze indices uit te rekenen een schatting van θ in de formule moeten invullen, maar deze schatting is een functie van de score, en met een bepaalde score *s* zijn niet alle mogelijke antwoordpatronen verdraagbaar. Indien bijvoorbeeld $s=1$, dan zijn er maar *k* verschillende antwoorden mogelijk bij deze score, en dus is het redelijk om alleen deze *k* antwoordpatronen te beoordelen op hun 'vreemdheid' onder het Raschmodel. Molenaar en Hoijtink hebben een index ontwikkeld waarbij dit ook gebeurt. De statistische significantie-toetsing van deze index is echter behoorlijk ingewikkeld.